Course:

**Optimization I – II**
**Convex Analysis**
**Nonlinear Programming Theory**
**Nonlinear Programming Algorithms**

♣ To make decisions optimally is one of the most basic desires of a human being.
Whenever the candidate decisions, design restrictions and design goals can be properly quantified, optimal decision-making yields an *optimization poblem*, most typically, a *Mathematical Programming* one:

$$
\begin{array}{lll}
\text{minimize} & & \\
 & f(x) & [\text{ objective }] \\
\text{subject to} & & \\
h_i(x) = 0,\ i=1,...,m & & \left[\begin{array}{l}\text{equality}\\ \text{constraints}\end{array}\right] \\
g_j(x) \leq 0,\ j=1,...,k & & \left[\begin{array}{l}\text{inequality}\\ \text{constraints}\end{array}\right] \\
x \in X & & [\text{ domain }]
\end{array}
\qquad \text{(MP)}
$$

♣ In (MP),
◇ a *solution* $x \in \mathbf{R}^n$ represents a candidate decision,
◇ the *constraints* express restrictions on the meaningful decisions (balance and state equations, bounds on recourses, etc.),
◇ the *objective* to be minimized represents the losses (minus profit) associated with a decision.

$$
\begin{array}{lll}
\text{minimize} & & \\
\qquad f(x) & & [\text{ objective }] \\
\text{subject to} & & \\
h_i(x) = 0,\ i = 1, ..., m & & \left[\begin{array}{l}\text{equality}\\ \text{constraints}\end{array}\right] \\
g_j(x) \leq 0,\ j = 1, ..., k & & \left[\begin{array}{l}\text{inequality}\\ \text{constraints}\end{array}\right] \\
\quad x \in X & & [\text{ domain }]
\end{array}
\qquad \text{(MP)}
$$

♣ To solve problem (MP) means to find its *optimal solution* $x_*$, that is, a *feasible* (i.e., satisfying the constraints) solution with the value of the objective $\leq$ its value at any other feasible solution:

$$
x_* : \left\{ \begin{array}{l} h_i(x_*) = 0\,\forall i \ \&\ g_j(x_*) \leq 0\,\forall j \ \&\ x_* \in X \\ h_i(x) = 0\,\forall i \ \&\ g_j(x) \leq 0\,\forall j \ \&\ x \in X \\ \qquad\qquad\qquad\qquad \Rightarrow f(x_*) \leq f(x) \end{array} \right.
$$

$$\min_x f(x)$$

$$\begin{array}{rcl} \text{s.t.} & & \\ h_i(x) & = & 0,\ i = 1, ..., m \\ g_j(x) & \leq & 0,\ j = 1, ..., k \\ x & \in & X \end{array} \qquad \text{(MP)}$$

♣ In *Combinatorial* (or *Discrete*) Optimization, the domain $X$ is a discrete set, like the set of all integral or 0/1 vectors.

In contrast to this, in *Continuous* Optimization we will focus on, $X$ is a "continuum" set like the entire $\mathbf{R}^n$,a *box* $\{x : a \leq x \leq b\}$, or *simplex* $\{x \geq 0 : \sum_j x_j = 1\}$, etc., and the objective and the constraints are (at least) continuous on $X$.

♣ In *Linear Programming*, $X = \mathbf{R}^n$ and the objective and the constraints are linear functions of $x$.
In contrast to this, *Nonlinear Continuous Optimization*, (some of) the objectives and the constraints are nonlinear.

$$
\begin{array}{rcl}
& \min\limits_x f(x) & \\
\text{s.t.} & & \\
h_i(x) & = & 0,\ i = 1, ..., m \\
g_j(x) & \leq & 0,\ j = 1, ..., k \\
x & \in & X
\end{array}
\qquad \text{(MP)}
$$

♣ The goals of our course is to present

- *basic theory* of Continuous Optimization, with emphasis on *existence* and *uniqueness* of optimal solutions and their *characterization* (i.e., necessary and/or sufficient optimality conditions);

- *traditional algorithms* for building (approximate) optimal solutions to Continuous Optimization problems.

♣ *Mathematical foundation* of Optimization Theory is given by *Convex Analysis* – a specific combination of of Real Analysis and Geometry unified by and focusing on investigating convexity-related notions.

## Theory of Systems of Linear Inequalities, I
## Homogeneous Farkas Lemma

♣ Question: *When a vector $a \in \mathbb{R}^n$ is a conic combination of given vectors $a_1, ..., a_m$ ?*

Answer: [Homogeneous Farkas Lemma] *A vector $a \in \mathbb{R}^n$ can be represented as a conic combination $\sum\limits_{i=1}^{m} \lambda_i a_i$, $\lambda_i \geq 0$, of given vectors $a_1, ..., a_m \in \mathbb{R}^n$ iff the homogeneous linear inequality*

$$a^T x \geq 0 \qquad (I)$$

*is a consequence of the system of homogeneous linear inequalities*

$$a_i^T x \geq 0,\ i = 1, ..., m \qquad (S)$$

*i.e., iff the following implication is true:*

$$a_i^T x \geq 0,\ i = 1, ..., m \Rightarrow a^T x \geq 0. \qquad (*)$$

**Proof, $\Rightarrow$:** If $a = \sum\limits_{i=1}^{m} \lambda_i a_i$ with $\lambda_i \geq 0$, then, of course, $a^T x = \sum\limits_{i} \lambda_i a_i^T x$ for all $x$, and $(*)$ clearly is true.

$$a_i^T x \geq 0,\, i = 1, ..., m \Rightarrow a^T x \geq 0. \qquad (*)$$

**Proof, $\Leftarrow$:** Assume that $(*)$ is true, and let us prove that $a$ is a conic combination of $a_1, ..., a_m$. The case of $a = 0$ is trivial, thus assume that $a \neq 0$.

**1**$^0$**.** Let

$$A_i = \left\{x : a^T x = -1, a_i^T x \geq 0\right\},\, i = 1, ..., m.$$

Note that $A_i$ are convex sets, and their intersection is empty by $(*)$. Consider a minimal, in number of sets, sub-family of the family $A_1, ..., A_m$ with empty intersection; w.l.o.g. we may assume that this sub-family is $A_1, ..., A_k$. Thus, the $k$ sets $A_1, ..., A_k$ do not intersect, while every $k-1$ sets from this family do intersect.

**2**$^0$**.** Claims: **A:** *Vector $a$ is a linear combination of vectors $a_1, ..., a_k$*

**B:** *Vectors $a_1, ..., a_k$ are linearly independent*

Situation: The sets $A_i = \{x : a^T x = -1, a_i^T x \geq 0\}$, $i = 1, ..., k$, are such that the $k$ sets $A_1, ..., A_k$ do not intersect, while every $k-1$ sets from this family do intersect.

Claim A: *Vector $a$ is a linear combination of vectors $a_1, ..., a_k$*

**Proof of A:** Assuming that $a \notin \mathrm{Lin}(\{a_1, ..., a_k\})$, there exists a vector $h$ with $a^T h \neq 0$ and $a_i^T h = 0$, $i = 1, ..., k$ (you can take as $h$ the projection of $a$ onto the orthogonal complement of $a_1, ..., a_k$). Setting $x = -(a^T h)^{-1} h$, we get $a^T x = -1$ and $a_i^T x = 0$, $i = 1, ..., k$, so that $x \in A_i$, $i = 1, ..., k$, which is a contradiction.

Situation: The sets $A_i = \{x : a^T x = -1, a_i^T x \geq 0\}$, $i = 1, ..., k$, are such that the $k$ sets $A_1, ..., A_k$ do not intersect, while every $k-1$ sets from this family do intersect.
Claim B: *Vectors $a_1, ..., a_k$ are linearly independent*

**Proof of B:** The case of $k = 1$ is evident. Indeed, in this case we should prove that $a_1 \neq 0$; if $a_1 = 0$ and $k = 1$, then by Claim A also $a = 0$, which is not the case. Thus, let us prove Claim B in the case of $k > 1$.

Assume, on the contrary, that

$$E = \text{Lin}(\{a_1, ..., a_k\})$$

is of dimension $r < k$, and let

$$B_i = \{x \in E : a^T x = -1, a_i^T x \geq 0\}.$$

$\diamondsuit$ By Claim A, $0 \neq a \in E$, so that $B_i$ belong to $(r-1)$-dimensional affine subspace $M = \{x \in E : a^T x = -1\}$. We claim that every $r$ of the sets $B_1, ..., B_k$ have a point in common. Indeed, to prove, e.g., that $B_1, ..., B_r$ have a point in common, note that $r < k$ and therefore the sets $A_1, ..., A_r$ have a point in common; projecting this point onto $E$, we clearly get a point from $B_1 \cap ... \cap B_r$.

$\diamondsuit$ $B_1, ..., B_k$ are convex subsets in $(m-1)$-dimensional affine subspace, and every $m$ of them have a common point. By Helley Theorem, all $k$ sets $B_1, ..., B_k$ have a common point, whence $A_1, ..., A_k$ have a common point, which is a contradiction.

$\mathbf{3^0.}$ By Claim A,

$$a = \sum_{i=1}^{k} \lambda_i a_i$$

with certain coefficients $\lambda_i$. All we need is to prove that $\lambda_i \geq 0$. To this end assume that certain $\lambda_i$, say, $\lambda_1$, is $< 0$. By claim B, the vectors $a_1, ..., a_k$ are linearly independent, and therefore there exists a vector $x$ such that

$$a_1^T x = 1,\ a_i^T x = 0,\ i = 2, ..., k.$$

Since $\lambda_1 < 0$, it follows that

$$a^T x = \lambda_1 a_1^T x < 0,$$

and the vector

$$\bar{x} = \frac{x}{|a^T x|}$$

clearly satisfies

$$a^T \bar{x} = -1,\ a_1^T \bar{x} > 0,\ a_i^T \bar{x} = 0,\ i = 2, ..., k,$$

that is, $\bar{x} \in A_1 \cap ... \cap A_k$, which is a contradiction.

## Theory of Systems of Linear Inequalities, II
## Theorem on Alternative

♣ A general (finite!) system of linear inequalities with unknowns $x \in \mathbb{R}^n$ can be written down as

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}} + 1, ..., m \end{array} \qquad (S)$$

Question: *How to certify that $(S)$ is solvable?*
Answer: A solution is a certificate of solvability!
Question: *How to certify that $S$ is not solvable?*
Answer: ???

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1,...,m_{\rm s} \\ a_i^T x & \geq & b_i,\ i = m_{\rm s}+1,...,m \end{array} \qquad (S)$$

Question: *How to certify that $S$ is not solvable?*

*Conceptual* sufficient insolvability condition:

*If we can lead the assumption that $x$ solves $(S)$ to a contradiction, then $(S)$ has no solutions.*

"Contradiction by linear aggregation": Let us associate with inequalities of $(S)$ *nonnegative* weights $\lambda_i$ and sum up the inequalities with these weights. The resulting inequality

$$\left[\sum_{i=1}^m \lambda_i a_i\right]^T x \left\{ \begin{array}{ll} > \sum\limits_i \lambda_i b_i, & \sum\limits_{i=1}^{m_{\rm s}} \lambda_s > 0 \\ \geq \sum\limits_i \lambda_i b_i, & \sum\limits_{i=1}^{m_{\rm s}} \lambda_s = 0 \end{array} \right. \qquad (C)$$

by its origin is a *consequence* of $(S)$, that is, it is satisfied at every solution to $(S)$.

Consequently, *if there exist $\lambda \geq 0$ such that $(C)$ has no solutions at all, then $(S)$ has no solutions!*

Question: When a linear inequality

$$d^T x \left\{ \begin{array}{c} > \\ \geq \end{array} \right. e$$

has no solutions at all?

Answer: This is the case if and only if $d = 0$ and

— either the sign is "$>$", and $e \geq 0$,

— or the sign is "$\geq$", and $e > 0$.

Conclusion: *Consider a system of linear inequalities*

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_s \\ a_i^T x & \geq & b_i,\ i = m_s + 1, ..., m \end{array} \qquad (S)$$

*in variables* $x$*, and let us associate with it two systems of linear inequalities in variables* $\lambda$*:*

$$\mathcal{T}_{\rm I}: \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum\limits_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum\limits_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{array} \right. \qquad \mathcal{T}_{\rm II}: \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum\limits_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum\limits_{i=1}^{m} \lambda_i b_i & > & 0 \end{array} \right.$$

*If one of the systems* $\mathcal{T}_{\rm I}$*,* $\mathcal{T}_{\rm II}$ *is solvable, then* $(S)$ *is unsolvable.*

Note: *If* $\mathcal{T}_{\rm II}$ *is solvable, then already the system*

$$a_i^T x \geq b_i,\ i = m_s + 1, ..., m$$

*is unsolvable!*

General Theorem on Alternative: *A system of linear inequalities*

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_s \\ a_i^T x & \geq & b_i,\ i = m_s + 1, ..., m \end{array} \qquad (S)$$

*is unsolvable iff one of the systems*

$$\mathcal{T}_{\mathrm{I}} : \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum\limits_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum\limits_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{array} \right. \qquad \mathcal{T}_{\mathrm{II}} : \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum\limits_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum\limits_{i=1}^{m} \lambda_i b_i & > & 0 \end{array} \right.$$

*is solvable.*

Note: *The subsystem*

$$a_i^T x \geq b_i,\ i = m_s + 1, ..., m$$

*of $(S)$ is unsolvable iff $\mathcal{T}_{\mathrm{II}}$ is solvable!*

**Proof.** We already know that solvability of one of the systems $\mathcal{T}_{\mathrm{I}}$, $\mathcal{T}_{\mathrm{II}}$ is a sufficient condition for unsolvability of $(S)$. All we need to prove is that if $(S)$ is unsolvable, then one of the systems $\mathcal{T}_{\mathrm{I}}$, $\mathcal{T}_{\mathrm{II}}$ is solvable.
Assume that the system

$$\begin{array}{lcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathrm{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathrm{s}} + 1, ..., m \end{array} \qquad (S)$$

in variables $x$ has no solutions. Then *every solution $x, \tau, \epsilon$ to the homogeneous system of inequalities*

$$\begin{array}{rrrcl} & \tau & -\epsilon & \geq & 0 \\ a_i^T x & -b_i\tau & -\epsilon & \geq & 0,\ i = 1, ..., m_{\mathrm{s}} \\ a_i^T x & -b_i\tau & & \geq & 0,\ i = m_{\mathrm{s}} + 1, ..., m \end{array}$$

*has $\epsilon \leq 0$.*
Indeed, in a solution with $\epsilon > 0$ one would also have $\tau > 0$, and the vector $\tau^{-1}x$ would solve $(S)$.

Situation: Every solution to the system of homogeneous inequalities

$$\begin{array}{rrrcl} & \tau & -\epsilon & \geq & 0 \\ a_i^T x & -b_i\tau & -\epsilon & \geq & 0,\ i=1,...,m_{\rm s} \\ a_i^T x & -b_i\tau & & \geq & 0,\ i=m_{\rm s}+1,...,m \end{array} \qquad (U)$$

has $\epsilon \leq 0$, i.e., the homogeneous inequality

$$-\epsilon \geq 0 \qquad (I)$$

is a consequence of system $(U)$ of homogeneous inequalities. By Homogeneous Farkas Lemma, *the vector of coefficients in the left hand side of $(I)$ is a conic combination of the vectors of coefficients in the left hand sides of $(U)$:*

$$\begin{array}{rcl} \exists \lambda \geq 0, \nu \geq 0 : & & \\ \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0 \\ -\sum\limits_{i=1}^{m} \lambda_i b_i + \nu & = & 0 \\ -\sum\limits_{i=1}^{m_{\rm s}} \lambda_i - \nu & = & -1 \end{array}$$

Assuming that $\lambda_1 = ... = \lambda_{m_{\rm s}} = 0$, we get $\nu = 1$, and therefore $\lambda$ solves $\mathcal{T}_{\rm II}$. In the case of $\sum\limits_{i=1}^{m_{\rm s}} \lambda_i > 0$, $\lambda$ clearly solves $\mathcal{T}_{\rm I}$.

# Corollaries of GTA

♣ Principle A: *A finite system of linear inequalities has no solutions iff one can lead it to a contradiction by linear aggregation, i.e., an appropriate weighted sum of the inequalities with "legitimate" weights is either a contradictory inequality*

$$0^T x > a \qquad [a \geq 0]$$

*or a contradictory inequality*

$$0^T x \geq a \qquad [a > 0]$$

♣ <u>Principle B:</u> [Inhomogeneous Farkas Lemma]

*A linear inequality*

$$a^T x \leq b$$

*is a consequence of <u>solvable</u> system of linear inequalities*

$$a_i^T x \leq b_i,\ i = 1, ..., m$$

*iff the target inequality can be obtained from the inequalities of the system <u>and</u> the identically true inequality*

$$0^T x \leq 1$$

*by linear aggregation, that is, iff there exist <u>nonnegative</u>* $\lambda_0, \lambda_1, ..., \lambda_m$ *such that*

$$\begin{array}{rcl} a & = & \sum\limits_{i=1}^m \lambda_i a_i \\ b & = & \lambda_0 + \sum\limits_{i=1}^m \lambda_i b_i \end{array} \quad \left[ \Leftrightarrow \begin{array}{rcl} a & = & \sum\limits_{i=1}^m \lambda_i a_i \\ b & \geq & \sum\limits_{i=1}^m \lambda_i b_i \end{array} \right]$$

Linear Programming Duality Theorem

♣ The origin of the LP dual of a Linear Programming program

$$\text{Opt}(P) = \min_x \left\{ c^T x : Ax \geq b \right\} \qquad (P)$$

is the desire to get a systematic way to bound from below the optimal value in $(P)$. The conceptually simplest bounding scheme is *linear aggregation of the constraints*:

Observation: *For every vector $\lambda$ of nonnegative weights, the constraint*

$$[A^T\lambda]^T x \equiv \lambda^T A x \geq \lambda^T b$$

*is a consequence of the constraints of $(P)$ and as such is satisfied at every feasible solution of $(P)$.*

Corollary: *For every vector $\lambda \geq 0$ such that $A^T\lambda = c$, the quantity $\lambda^T b$ is a lower bound on $\text{Opt}(P)$.*

♣ *The problem dual to $(P)$ is nothing but the problem*

$$\text{Opt}(D) = \max_\lambda \left\{ b^T\lambda : \lambda \geq 0, A^T\lambda = c \right\} \quad (D)$$

*of maximizing the lower bound on $\text{Opt}(P)$ given by Corollary.*

♣ The origin of $(D)$ implies the following Weak Duality Theorem: The value of the primal objective at every feasible solution of the primal problem

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : Ax \geq b \right\} \qquad (P)$$

is $\geq$ the value of the dual objective at every feasible solution to the dual problem

$$\mathrm{Opt}(D) = \max_\lambda \left\{ b^T \lambda : \lambda \geq 0, A^T \lambda = c \right\} \quad (D)$$

that is,

$$\left.\begin{array}{l} x \text{ is feasible for } (P) \\ \lambda \text{ is feasible for } (D) \end{array}\right\} \Rightarrow c^T x \geq b^T \lambda$$

In particular,

$$\mathrm{Opt}(P) \geq \mathrm{Opt}(D).$$

♣ LP Duality Theorem: *Consider an LP program along with its dual:*

$$
\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min\limits_{x}\left\{c^T x : Ax \geq b\right\} & (P) \\
\mathrm{Opt}(D) & = & \max\limits_{\lambda}\left\{b^T\lambda : A^T\lambda = c, \lambda \geq 0\right\} & (D)
\end{array}
$$

*Then*

♢ *Duality is symmetric: the problem dual to dual is (equivalent to) the primal*

♢ *The value of the dual objective at every dual feasible solution is $\leq$ the value of the primal objective at every primal feasible solution*

♢ *The following 5 properties are equivalent to each other:*

- *(i)* $(P)$ *is feasible and bounded (below)*
- *(ii)* $(D)$ *is feasible and bounded (above)*
- *(iii)* $(P)$ *is solvable*
- *(iv)* $(D)$ *is solvable*
- *(v)* *both* $(P)$ *and* $(D)$ *are feasible*

*and whenever they take place, one has* $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_x \left\{c^T x : Ax \geq b\right\} & (P) \\ \mathrm{Opt}(D) & = & \max\limits_\lambda \left\{b^T \lambda : A^T \lambda = c, \lambda \geq 0\right\} & (D) \end{array}$$

♢ *Duality is symmetric*

**Proof:** Rewriting $(D)$ in the form of $(P)$, we arrive at the problem

$$\min_\lambda \left\{ -b^T \lambda : \left[\begin{array}{c} A^T \\ -A^T \\ I \end{array}\right] \lambda \geq \left[\begin{array}{c} c \\ -c \\ 0 \end{array}\right] \right\},$$

with the dual being

$$\max_{u,v,w} \left\{ c^T u - c^T v + 0^T w : \begin{array}{c} u \geq 0, v \geq 0, w \geq 0, \\ Au - Av + w = -b \end{array} \right\}$$

$$\Updownarrow$$

$$\max_{x = v-u, w} \left\{ -c^T x : w \geq 0, Ax = b + w \right\}$$

$$\Updownarrow$$

$$\min_x \left\{ c^T x : Ax \geq b \right\}$$

$\diamondsuit$ The value of the dual objective at every dual feasible solution is $\leq$ the value of the primal objective at every primal feasible solution

This is Weak Duality

$\diamondsuit$ The following 5 properties are equivalent to each other:

$$\begin{array}{cr} (P) \text{ is feasible and bounded below} & \text{(i)} \\ \Downarrow & \\ (D) \text{ is solvable} & \text{(iv)} \end{array}$$

Indeed, by origin of $\mathrm{Opt}(P)$, the inequality

$$c^T x \geq \mathrm{Opt}(P)$$

is a consequence of the (solvable!) system of inequalities

$$Ax \geq b.$$

By Principle B, the inequality is a linear consequence of the system:

$$\exists \lambda \geq 0 : A^T\lambda = c \ \& \ b^T\lambda \geq \mathrm{Opt}(P).$$

Thus, the dual problem has a feasible solution with the value of the dual objective $\geq \mathrm{Opt}(P)$. By Weak Duality, this solution is optimal, and $\mathrm{Opt}(D) = \mathrm{Opt}(P)$.

$\diamondsuit$ The following 5 properties are equivalent to each other:

$$
\begin{array}{c}
(D) \text{ is solvable} \\
\Downarrow \\
(D) \text{ is feasible and bounded above}
\end{array}
\quad
\begin{array}{c}
\text{(iv)} \\
\\
\text{(ii)}
\end{array}
$$

Evident

$\diamondsuit$ The following 5 properties are equivalent to each other:

$(D)$ is feasible and bounded above (ii)

$\Downarrow$

$(P)$ is solvable (iii)

Implied by already proved relation

$(P)$ is feasible and bounded below (i)

$\Downarrow$

$(D)$ is solvable (iv)

in view of primal-dual symmetry

$\diamondsuit$ The following 5 properties are equivalent to each other:

$$(P) \text{ is solvable} \qquad \text{(iii)}$$
$$\Downarrow$$
$$(P) \text{ is feasible and bounded below} \qquad \text{(i)}$$

Evident

We proved that

$$\text{(i)} \Leftrightarrow \text{(ii)} \Leftrightarrow \text{(iii)} \Leftrightarrow \text{(iv)}$$

and that when these 4 equivalent properties take place, one has

$$\text{Opt}(P) = \text{Opt}(D)$$

It remains to prove that properties (i) − (iv) are equivalent to

$$\text{both } (P) \text{ and } (D) \text{ are feasible} \qquad \text{(v)}$$

♢ In the case of (v), $(P)$ is feasible and below bounded (Weak Duality), so that (v)⇒(i)
♢ in the case of (i)≡(ii), both $(P)$ and $(D)$ are feasible, so that (i)⇒(v)

## Optimality Conditions in LP

Theorem: *Consider a primal-dual pair of feasible LP programs*

$$
\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min\limits_{x} \left\{ c^T x : Ax \geq b \right\} & (P) \\
\mathrm{Opt}(D) & = & \max\limits_{\lambda} \left\{ b^T \lambda : A^T \lambda = c, \lambda \geq 0 \right\} & (D)
\end{array}
$$

*and let $x, \lambda$ be feasible solutions to the respective programs. These solutions are optimal for the respective problems*

♢ *iff* $c^T x - b^T \lambda = 0$ ["zero duality gap"]

*as well as*

♢ *iff* $[Ax - b]_i \cdot \lambda_i = 0$ *for all* $i$ ["complementary slackness"]

**Proof:** Under Theorem's premise, $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, so that

$$
c^T x - b^T \lambda = \underbrace{c^T x - \mathrm{Opt}(P)}_{\geq 0} + \underbrace{\mathrm{Opt}(D) - b^T \lambda}_{\geq 0}
$$

Thus, duality gap $c^T x - b^T \lambda$ is always nonnegative and is zero iff $x, \lambda$ are optimal for the respective problems.

The complementary slackness condition is given by the identity

$$c^T x - b^T \lambda = (A^T \lambda)^T x - b^T \lambda = [Ax - b]^T \lambda$$

Since both $[Ax-b]$ and $\lambda$ are nonnegative, duality gap is zero iff the complementary slackness holds true.

# Separation Theorem

♣ Every linear form $f(x)$ on $\mathbb{R}^n$ is representable via inner product:

$$f(x) = f^T x$$

for appropriate vector $f \in \mathbb{R}^n$ uniquely defined by the form. Nontrivial (not identically zero) forms correspond to nonzero vectors $f$.

♣ A *level set*

$$M = \left\{ x : f^T x = a \right\} \qquad (*)$$

of a *nontrivial* linear form on $\mathbb{R}^n$ is affine subspace of affine dimension $n-1$; vice versa, every affine subspace $M$ of affine dimension $n-1$ in $\mathbb{R}^n$ can be represented by $(*)$ with appropriately chosen $f \neq 0$ and $a$; $f$ and $a$ are defined by $M$ up to multiplication by a common nonzero factor.

$(n-1)$-dimensional affine subspaces in $\mathbb{R}^n$ are called *hyperplanes*.

$$M = \left\{ x : f^T x = a \right\} \qquad (*)$$

♣ Level set $(*)$ of nontrivial linear form splits $\mathbb{R}^n$ into two parts:

$$\begin{array}{rcl} M_+ & = & \{x : f^T x \geq a\} \\ M_- & = & \{x : f^T x \leq a\} \end{array}$$

called *closed half-spaces* given by $(f, a)$; the hyperplane $M$ is the common boundary of these half-spaces. The interiors $M_{++}$ of $M_+$ and $M_{--}$ of $M_-$ are given by

$$\begin{array}{rcl} M_{++} & = & \{x : f^T x > a\} \\ M_{--} & = & \{x : f^T x < a\} \end{array}$$

and are called *open half-spaces* given by $(f, a)$. We have

$$\mathbb{R}^n = M_- \bigcup M_+ \quad [M_- \bigcap M_+ = M]$$

and

$$\mathbb{R}^n = M_{--} \bigcup M \bigcup M_{++}$$

♣ Definition. Let $T, S$ be two nonempty sets in $\mathbb{R}^n$.

(i) We say that a hyperplane

$$M = \{x : f^T x = a\} \qquad (*)$$

*separates* $S$ and $T$, if

♢ $S \subset M_-$, $T \subset M_+$ ("$S$ does not go above $M$, and $T$ does not go below $M$")

and

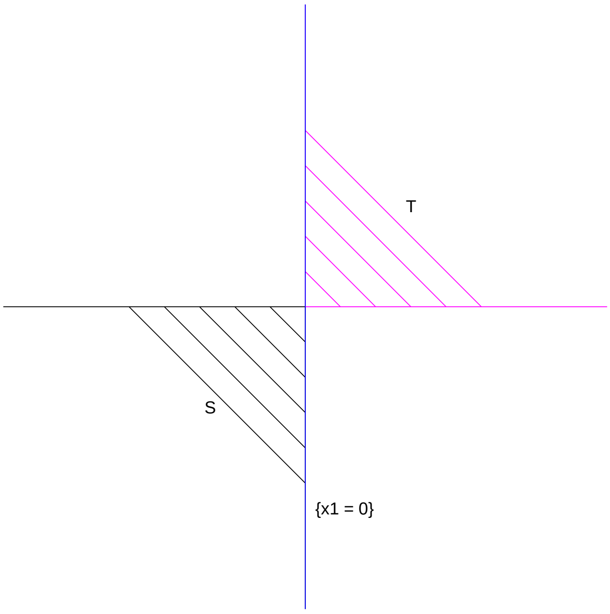
♢ $S \cup T \not\subset M$.

(ii) We say that a nontrivial linear form $f^T x$ separates $S$ and $T$ if, for properly chosen $a$, the hyperplane $(*)$ separates $S$ and $T$.

Examples: The linear form $x_1$ on $\mathbb{R}^2$

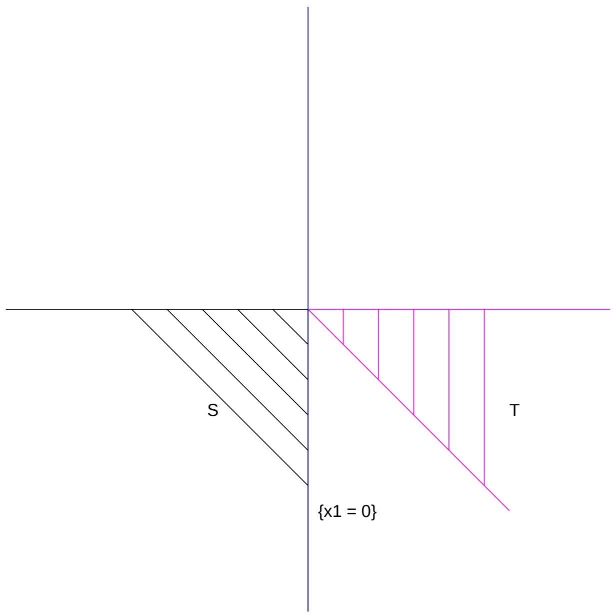
1) separates the sets

$$\begin{array}{rcl} S &=& \{x \in \mathbb{R}^2 : x_1 \leq 0, x_2 \leq 0\}, \\ T &=& \{x \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\} : \end{array}$$



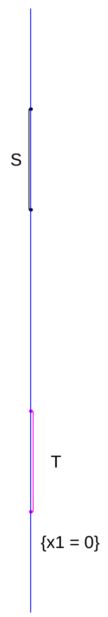
2) separates the sets

$$
\begin{aligned}
S &= \{x \in \mathbb{R}^2 : x_1 \le 0, x_2 \le 0\}, \\
T &= \{x \in \mathbb{R}^2 : x_1 + x_2 \ge 0, x_2 \le 0\} :
\end{aligned}
$$



3) does <u>not</u> separate the sets

$$
\begin{aligned}
S &= \{x \in \mathbb{R}^2 : x_1 = 0, 1 \leq x_2 \leq 2\}, \\
T &= \{x \in \mathbb{R}^2 : x_1 = 0, -2 \leq x_2 \leq -1\} :
\end{aligned}
$$



Observation: *A linear form $f^T x$ separates nonempty sets $S$, $T$ iff*

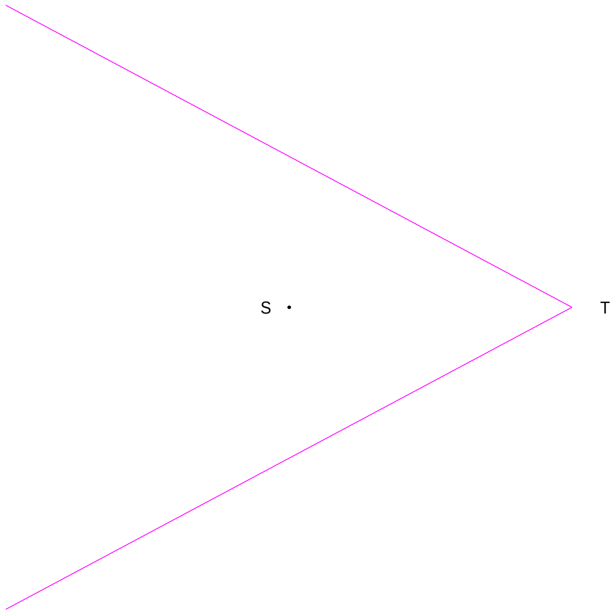
$$\begin{array}{rcl} \sup\limits_{x\in S} f^T x & \leq & \inf\limits_{y\in T} f^T y \\ \inf\limits_{x\in S} f^T x & < & \sup\limits_{y\in T} f^T y \end{array} \qquad (*)$$

*In the case of $(*)$, the associated with $f$ hyperplanes separating $S$ and $T$ are exactly the hyperplanes*

$$\{x : f^T x = a\} \text{ with } \sup_{x\in S} f^T x \leq a \leq \inf_{y\in T} f^T y.$$

♣ Separation Theorem: *Two nonempty convex sets $S$, $T$ can be separated iff their relative interiors do not intersect.*

Note: In this statement, convexity of both $S$ and $T$ is crucial!



**Proof, $\Rightarrow$:** *(!) <u>If</u> nonempty convex sets $S$, $T$ can be separated, <u>then</u>* $\mathrm{rint}\, S \bigcap \mathrm{rint}\, T = \emptyset$

<u>Lemma.</u> *Let $X$ be a convex set, $f(x) = f^T x$ be a linear form and $a \in \mathrm{rint}\, X$. Then*

$$f^T a = \max_{x \in X} f^T x \Leftrightarrow f(\cdot)\Big|_X = \mathrm{const}.$$

♣ <u>Lemma $\Rightarrow$ (!):</u> Let $a \in \mathrm{rint}\, S \cap \mathrm{rint}\, T$. Assume, on contrary to what should be proved, that $f^T x$ separates $S$, $T$, so that

$$\sup_{x \in S} f^T x \le \inf_{y \in T} f^T y.$$

♢ Since $a \in T$, we get $f^T a \ge \sup_{x \in S} f^T x$, that is, $f^T a = \max_{x \in S} f^T x$. By Lemma, $f^T x = f^T a$ for all $x \in S$.

♢ Since $a \in S$, we get $f^T a \le \inf_{y \in T} f^T y$, that is, $f^T a = \min_{y \in T} f^T y$. By Lemma, $f^T y = f^T a$ for all $y \in T$.

Thus,

$$z \in S \cup T \Rightarrow f^T z \equiv f^T a,$$

so that $f$ does <u>not</u> separate $S$ and $T$, which is a contradiction.

<u>Lemma.</u> *Let $X$ be a convex set, $f(x) = f^T x$ be a linear form and $a \in \operatorname{rint} X$. Then*

$$f^T a = \max_{x \in X} f^T x \Leftrightarrow f(\cdot)\Big|_X = \text{const.}$$

**Proof.** Shifting $X$, we may assume $a = 0$. Let, on the contrary to what should be proved, $f^T x$ be non-constant on $X$, so that there exists $y \in X$ with $f^T y \neq f^T a = 0$. The case of $f^T y > 0$ is impossible, since $f^T a = 0$ is the maximum of $f^T x$ on $X$. Thus, $f^T y < 0$. The line $\{ty : t \in \mathbb{R}\}$ passing through 0 and through $y$ belongs to $\mathrm{Aff}(X)$; since $0 \in \operatorname{rint} X$, all points $z = -\epsilon y$ on this line belong to $X$, provided that $\epsilon > 0$ is small enough. At every point of this type, $f^T z > 0$, which contradicts the fact that $\max\limits_{x \in X} f^T x = f^T a = 0$.

**Proof, $\Leftarrow$:** Assume that $S$, $T$ are nonempty convex sets such that $\mathrm{rint}\, S \cap \mathrm{rint}\, T = \emptyset$, and let us prove that $S$, $T$ can be separated.

**Step 1: Separating a point and a convex hull of a finite set.** Let $S = \mathrm{Conv}(\{b_1, ..., b_m\})$ and $T = \{b\}$ with $b \notin S$, and let us prove that $S$ and $T$ can be separated.

**$1^0$.** Let

$$\beta_i = \begin{bmatrix} b_i \\ 1 \end{bmatrix}, \ \beta = \begin{bmatrix} b \\ 1 \end{bmatrix}.$$

Observe that $\beta$ is *not* a conic combination of $\beta_1, ..., \beta_m$:

$$\begin{bmatrix} b \\ 1 \end{bmatrix} = \sum_{i=1}^{m} \lambda_i \begin{bmatrix} b_i \\ 1 \end{bmatrix}, \lambda_i \geq 0$$

$$\Downarrow$$

$$b = \sum_i \lambda_i b_i, \ \sum_i \lambda_i = 1, \ \lambda_i \geq 0$$

$$\Downarrow$$

$$b \in S - \text{contradiction!}$$

$$\beta_i = \begin{bmatrix} b_i \\ 1 \end{bmatrix}, \ \beta = \begin{bmatrix} b \\ 1 \end{bmatrix}.$$

**$2^0$.** Since $\beta$ is not conic combination of $\beta_i$, by Homogeneous Farkas Lemma there exists $h = \begin{bmatrix} f \\ a \end{bmatrix}$ such that

$$f^T b - a \equiv h^T \beta > 0 \geq h^T \beta_i \equiv f^T b_i - a, \ i = 1, ..., m$$

that is,

$$f^T b > \max_{i=1,...,m} f^T b_i = \max_{x \in S = \mathrm{Conv}(\{b_1,...,b_m\})} f^T x.$$

Note: We have used the evident fact that

$$\begin{aligned} \max_{x \in \mathrm{Conv}(\{b_1,...,b_m\})} f^T x &\equiv \max_{\lambda \geq 0, \sum_i \lambda_i = 1} f^T [\sum_i \lambda_i b_i] \\ &= \max_{\lambda \geq 0, \sum_i \lambda_i = 1} \sum_i \lambda_i [f^T b_i] \\ &= \max_i f^T b_i. \end{aligned}$$

**Step 2: Separating a point and a convex set which does not contain the point.** Let $S$ be a nonempty convex set and $T = \{b\}$ with $b \notin S$, and let us prove that $S$ and $T$ can be separated.

**$1^0$.** Shifting $S$ and $T$ by $-b$ (which clearly does not affect the possibility of separating the sets), we can assume that $T = \{0\} \not\subset S$.

**$2^0$.** Replacing, if necessary, $\mathbb{R}^n$ with Lin($S$), we may further assume that $\mathbb{R}^n = $ Lin($S$).

Lemma: *Every nonempty subset $S$ in $\mathbf{R}^n$ is separable: one can find a sequence $\{x_i\}$ of points from $S$ which is dense in $S$, i.e., is such that every point $x \in S$ is the limit of an appropriate subsequence of the sequence.*

**Lemma $\Rightarrow$ Separation:** Let $\{x_i \in S\}$ be a sequence which is dense in $S$. Since $S$ is convex and does not contain 0, we have

$$0 \notin \mathrm{Conv}(\{x_1, ..., x_i\}) \ \forall i$$

whence

$$\exists f_i : 0 = f_i^T 0 > \max_{1 \le j \le i} f_i^T x_j. \qquad (*)$$

By scaling, we may assume that $\|f_i\|_2 = 1$. The sequence $\{f_i\}$ of unit vectors possesses a converging subsequence $\{f_{i_s}\}_{s=1}^{\infty}$; the limit $f$ of this subsequence is, of course, a unit vector. By $(*)$, for every fixed $j$ and all large enough $s$ we have $f_{i_s}^T x_j < 0$, whence

$$f^T x_j \le 0 \ \forall j. \qquad (**)$$

Since $\{x_j\}$ is dense in $S$, $(**)$ implies that $f^T x \le 0$ for all $x \in S$, whence

$$\sup_{x \in S} f^T x \le 0 = f^T 0.$$

Situation: (a) $\text{Lin}(S) = \mathbb{R}^n$
(b) $T = \{0\}$
(c) We have built a unit vector $f$ such that

$$\sup_{x \in S} f^T x \leq 0 = f^T 0. \qquad (!)$$

By (!), all we need to prove that $f$ separates $T = \{0\}$ and $S$ is to verify that

$$\inf_{x \in S} f^T x < f^T 0 = 0.$$

Assuming the opposite, (!) would say that $f^T x = 0$ for all $x \in S$, which is impossible, since $\text{Lin}(S) = \mathbb{R}^n$ and $f$ is nonzero.

Lemma: *Every nonempty subset $S$ in $\mathbf{R}^n$ is separable: one can find a sequence $\{x_i\}$ of points from $S$ which is dense in $S$, i.e., is such that every point $x \in S$ is the limit of an appropriate subsequence of the sequence.*

**Proof.** Let $r_1, r_2, ...$ be the countable set of all rational vectors in $\mathbb{R}^n$. For every positive integer $t$, let $X_t \subset S$ be the countable set given by the following construction:

> We look, one after another, at the points $r_1, r_2, ...$ and for every point $r_s$ check whether there is a point $z$ in $S$ which is at most at the distance $1/t$ away from $r_s$. If points $z$ with this property exist, we take one of them and add it to $X_t$ and then pass to $r_{s+1}$, otherwise directly pass to $r_{s+1}$.

Is is clear that

> (*) Every point $x \in S$ is at the distance at most $2/t$ from certain point of $X_t$.

Indeed, since the rational vectors are dense in $\mathbb{R}^n$, there exists $s$ such that $r_s$ is at the distance $\leq \frac{1}{t}$ from $x$. Therefore, when processing $r_s$, we definitely add to $X_t$ a point $z$ which is at the distance $\leq 1/t$ from $r_s$ and thus is at the distance $\leq 2/t$ from $x$.
By construction, the countable union $\bigcup\limits_{t=1}^{\infty} X_t$ of countable sets $X_t \subset S$ is a countable set in $S$, and by (*) this set is dense in $S$.

**Step 3: Separating two non-intersecting nonempty convex sets.** Let $S$, $T$ be nonempty convex sets which do not intersect; let us prove that $S, T$ can be separated. Let $\widehat{S} = S - T$ and $\widehat{T} = \{0\}$. The set $\widehat{S}$ clearly is convex and does not contain 0 (since $S \cap T = \emptyset$). By Step 2, $\widehat{S}$ and $\{0\} = \widehat{T}$ can be separated: there exists $f$ such that

$$\left\{\begin{array}{lcl} \overbrace{\sup\limits_{x\in S, y\in T}[f^Tx - f^Ty]}^{\sup\limits_{x\in S} f^Ts - \inf\limits_{y\in T} f^Ty} & \leq & 0 = \inf\limits_{z\in\{0\}} f^Tz \\ \underbrace{\inf\limits_{x\in S, y\in T}[f^Tx - f^Ty]}_{\inf\limits_{x\in S} f^Tx - \sup\limits_{y\in T} f^Ty} & < & 0 = \sup\limits_{z\in\{0\}} f^Tz \end{array}\right.$$

whence

$$\begin{array}{rcl} \sup\limits_{x\in S} f^Tx & \leq & \inf\limits_{y\in T} f^Ty \\ \inf\limits_{x\in S} f^Tx & < & \sup\limits_{y\in T} f^Ty \end{array}$$

**Step 4: Completing the proof of Separation Theorem.** Finally, let $S$, $T$ be nonempty convex sets with non-intersecting relative interiors, and let us prove that $S$, $T$ can be separated.

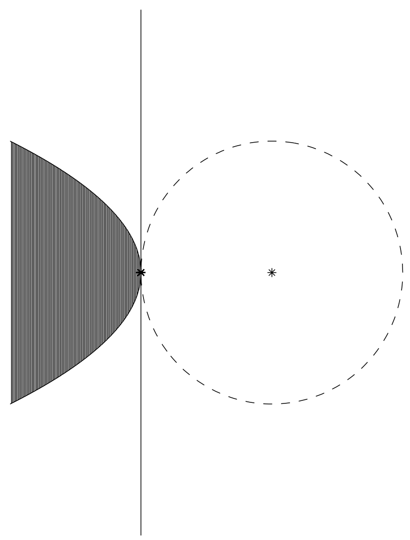
As we know, the sets $S' = \mathrm{rint}\, S$ and $T' = \mathrm{rint}\, T$ are convex and nonempty; we are in the situation when these sets do not intersect. By Step 3, $S'$ and $T'$ can be separated: for properly chosen $f$, one has

$$\begin{array}{rcl} \sup\limits_{x\in S'} f^T x & \leq & \inf\limits_{y\in T'} f^T y \\ \inf\limits_{x\in S'} f^T x & < & \sup\limits_{y\in T'} f^T y \end{array} \qquad (*)$$

Since $S'$ is dense in $S$ and $T'$ is dense in $T$, inf's and sup's in $(*)$ remain the same when replacing $S'$ with $S$ and $T'$ with $T$. Thus, $f$ separates $S$ and $T$.

♣ Alternative proof of Separation Theorem starts with separating a point $T = \{a\}$ and a *closed* convex set $S$, $a \notin S$, and is based on the following fact:

> *Let $S$ be a nonempty closed convex set and let $a \notin S$. There exists a unique closest to $a$ point in $S$:*
>
> $$\mathrm{Proj}_S(a) = \operatorname*{argmin}_{x \in S} \|a - x\|_2$$
>
> *and the vector $e = a - \mathrm{Proj}_S(a)$ separates $a$ and $S$:*
>
> $$\max_{x \in S} e^T x = e^T \mathrm{Proj}_S(a) = e^T a - \|e\|_2^2 < e^T a.$$



**Proof: $1^0$.** The closest to $a$ point in $S$ does exist. Indeed, let $x_i \in S$ be a sequence such that

$$\|a - x_i\|_2 \to \inf_{x \in S} \|a - x\|_2, \ , \ i \to \infty$$

The sequence $\{x_i\}$ clearly is bounded; passing to a subsequence, we may assume that $x_i \to \bar{}$ as $i \to \infty$. *Since $S$ is closed*, we have $\bar{x} \in S$, and

$$\|a - \bar{x}\|_2 = \lim_{i \to \infty} \|a - x_i\|_2 = \inf_{x \in S} \|a - x\|_2.$$

**$2^0$.** The closest to $a$ point in $S$ is unique. Indeed, let $x, y$ be two closest to $a$ points in $S$, so that $\|a - x\|_2 = \|a - y\|_2 = d$. *Since $S$ is convex*, the point $z = \frac{1}{2}(x + y)$ belongs to $S$; therefore $\|a - z\|_2 \geq d$. We now have

$$\begin{aligned} &\overbrace{\|[a-x]+[a-y]\|_2^2}^{=\|2(a-z)\|_2^2 \geq 4d^2} + \overbrace{\|[a-x]-[a-y]\|_2^2}^{=\|x-y\|^2} \\ &= \underbrace{2\|a-x\|_2^2 + 2\|a-y\|_2^2}_{4d^2} \end{aligned}$$

whence $\|x - y\|_2 = 0$.

**$3^0$.** Thus, the closest to $a$ point in $S$ exists and is unique. With $e = a - \mathrm{Proj}_S(a)$, we have

$$
\begin{array}{rcl}
& x \in S, f = x - \mathrm{Proj}_S(a) & \\
& \Downarrow & \\
\phi(t) & \equiv & \|e - tf\|_2^2 \\
& = & \|a - [\mathrm{Proj}_S(a) + t(x - \mathrm{Proj}_S(a))]\|_2^2 \\
& \geq & \|a - \mathrm{Proj}_S(a)\|_2^2 \\
& = & \phi(0), 0 \leq t \leq 1 \\
& \Downarrow & \\
& 0 \leq \phi'(0) = -2e^T(x - \mathrm{Proj}_S(a)) & \\
& \Downarrow & \\
& \forall x \in S : e^T x \leq e^T \mathrm{Proj}_S(a) = e^T a - \|e\|_2^2. &
\end{array}
$$

♣ Separation of sets $S, T$ by linear form $f^T x$ is called *strict*, if

$$\sup_{x \in S} f^T x < \inf_{y \in T} f^T y$$

Theorem: Let $S, T$ be nonempty convex sets. These sets can be strictly separated iff they are at positive distance:

$$\text{dist}(S, T) = \inf_{x \in S, y \in T} \|x - y\|_2 > 0.$$

**Proof, $\Rightarrow$:** Let $f$ strictly separate $S, T$; let us prove that $S, T$ are at positive distance. Otherwise we could find sequences $x_i \in S$, $y_i \in T$ with $\|x_i - y_i\|_2 \to 0$ as $i \to \infty$, whence $f^T(y_i - x_i) \to 0$ as $i \to \infty$. It follows that the sets on the axis

$$\widehat{S} = \{a = f^T x : x \in S\}, \widehat{T} = \{b = f^T y : y \in T\}$$

are at zero distance, which is a contradiction with

$$\sup_{a \in \widehat{S}} a < \inf_{b \in \widehat{T}} b.$$

**Proof, $\Leftarrow$:** Let $T$, $S$ be nonempty convex sets which are at positive distance $2\delta$:

$$2\delta = \inf_{x\in S, y\in T} \|x-y\|_2 > 0.$$

Let

$$S^+ = S + \{z : \|z\|_2 \le \delta\}$$

The sets $S^+$ and $T$ are convex and do not intersect, and thus can be separated:

$$\sup_{x_+\in S^+} f^T x_+ \le \inf_{y\in T} f^T y \qquad [f \neq 0]$$

Since

$$\begin{array}{rcl}\sup\limits_{x_+\in S^+} f^T x_+ & = & \sup\limits_{x\in S, \|z\|_2\le\delta} [f^T x + f^T z] \\ & = & [\sup\limits_{x\in S} f^T x] + \delta\|f\|_2,\end{array}$$

we arrive at

$$\sup_{x\in S} f^T x < \inf_{y\in T} f^T y$$

Exercise Below $S$ is a nonempty convex set and $T = \{a\}$.

| Statement | True? |
|---|---|
| If $T$ and $S$ can be separated then $a \notin S$ | |
| If $a \notin S$, then $T$ and $S$ can be separated | |
| If $T$ and $S$ can be strictly separated, then $a \notin S$ | |
| If $a \notin S$, then $T$ and $S$ can be strictly separated | |
| If $S$ is closed and $a \notin S$, then $T$ and $S$ can be strictly separated | |

## Supporting Planes and Extreme Points

♣ Definition. Let $Q$ be a *closed* convex set in $\mathbb{R}^n$ and $\bar{x}$ be a point from the relative boundary of $Q$. A hyperplane

$$\Pi = \{x : f^T x = a\} \qquad [a \neq 0]$$

is called *supporting to $Q$ at the point $\bar{x}$*, if the hyperplane separates $Q$ and $\{\bar{x}\}$:

$$\begin{array}{rcl} \sup\limits_{x \in Q} f^T x & \leq & f^T \bar{x} \\ \inf\limits_{x \in Q} f^T x & < & f^t \bar{x} \end{array}$$

Equivalently: Hyperplane $\Pi = \{x : f^T x = a\}$ supports $Q$ at $\bar{x}$ iff the linear form $f^T x$ attains its maximum on $Q$, equal to $a$, at the point $\bar{x}$ and the form is non-constant on $Q$.

Proposition. Let $Q$ be a convex closed set in $\mathbb{R}^n$ and $\bar{x}$ be a point from the relative boundary of $Q$. Then

$\diamondsuit$ There exist at least one hyperplane $\Pi$ which supports $Q$ at $\bar{x}$;

$\diamondsuit$ For every such hyperplane $\Pi$, the set $Q \cap \Pi$ has dimension less than the one of $Q$.

**Proof:** Existence of supporting plane is given by Separation Theorem. This theorem is applicable since

$$\bar{x} \notin \mathrm{rint}\, Q \Rightarrow \{\bar{x}\} \equiv \mathrm{rint}\, \{\bar{x}\} \cap \mathrm{rint}\, Q = \emptyset.$$

Further,

$$Q \not\subseteq \Pi \Rightarrow \mathrm{Aff}(Q) \not\subseteq \Pi \Rightarrow \mathrm{Aff}(\Pi \cap Q) \subsetneqq \mathrm{Aff}(Q),$$

and if two *distinct* affine subspaces are embedded one into another, then the dimension of the embedded subspace is strictly less than the dimension of the embedding one.

Extreme Points

♣ Definition. Let $Q$ be a convex set in $\mathbb{R}^n$ and $\bar{x}$ be a point of $X$. The point is called *extreme*, if it is not a convex combination, with positive weights, of two points of $X$ distinct from $\bar{x}$:

$$\bar{x} \in \mathrm{Ext}(Q)$$
$$\Updownarrow$$
$$\left\{ \{\bar{x} \in Q\} \;\; \& \;\; \left\{ \begin{array}{c} u, v \in Q, \lambda \in (0,1) \\ \bar{x} = \lambda u + (1-\lambda)v \end{array} \right\} \Rightarrow u = v = \bar{x} \right\}$$

Equivalently: A point $\bar{x} \in Q$ is extreme iff it is not the midpoint of a nontrivial segment in $Q$:

$$x \pm h \in Q \Rightarrow h = 0.$$

Equivalently: A point $\bar{x} \in Q$ is extreme iff the set $Q \backslash \{\bar{x}\}$ is convex.

Examples:

1. Extreme points of $[x, y]$ are ...

2. Extreme points of $\triangle ABC$ are ...

3. Extreme points of the ball $\{x : \|x\|_2 \leq 1\}$ are ...

Theorem [Krein-Milman] Let $Q$ be a closed convex and nonempty set in $\mathbb{R}^n$. Then

$\diamondsuit$ $Q$ possess extreme points iff $Q$ does not contain lines;

$\diamondsuit$ If $Q$ is bounded, then $Q$ is the convex hull of its extreme points:

$$Q = \mathrm{Conv}(\mathrm{Ext}(Q))$$

so that every point of $Q$ is convex combination of extreme points of $Q$.

Note: If $Q = \mathrm{Conv}(A)$, then $\mathrm{Ext}(Q) \subset A$. Thus, extreme points of a *closed convex bounded* set $Q$ give the *minimal* representation of $Q$ as Conv(...).

**Proof. $1^0$:** If closed convex set $Q$ does not contain lines, then $\mathrm{Ext}(Q) \neq \emptyset$

Important lemma: Let $S$ be a closed convex set and $\Pi = \{x : f^T x = a\}$ be a hyperplane which supports $S$ at certain point. Then

$$\mathrm{Ext}(\Pi \cap S) \subset \mathrm{Ext}(S).$$

**Proof of Lemma.** Let $\bar{x} \in \mathrm{Ext}(\Pi \cap S)$; we should prove that $\bar{x} \in \mathrm{Ext}(S)$. Assume, on the contrary, that $\bar{x}$ is a midpoint of a nontrivial segment $[u, v] \subset S$. Then $f^T \bar{x} = a = \max\limits_{x \in S} f^T x$, whence $f^T \bar{x} = \max\limits_{x \in [u,v]} f^T x$. A linear form can attain its maximum on a segment at the midpoint of the segment iff the form is constant on the segment; thus, $a = f^T \bar{x} = f^T u = f^T v$, that is, $[u, v] \subset \Pi \cap S$. But $\bar{x}$ is an extreme point of $\Pi \cap S$ – contradiction!

Let $Q$ be a nonempty closed convex set which does not contain lines. In order to build an extreme point of $Q$, apply the *Purification algorithm*:

Initialization: Set $S_0 = Q$ and choose $x_0 \in Q$.

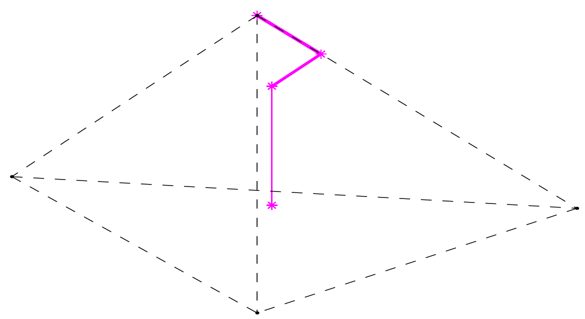
Step $t$: Given a nonempty closed convex set $S_t$ which does not contain lines and is such that $\mathrm{Ext}(S_t) \subset \mathrm{Ext}(Q)$ and $x_t \in S_t$,

1) check whether $S_t$ is a singleton $\{x_t\}$. If it is the case, terminate: $x_t \in \mathrm{Ext}\{S_t\} \subset \mathrm{Ext}(Q)$.

2) if $S_t$ is not a singleton, find a point $x_{t+1}$ on the relative boundary of $S_t$ and build a hyperplane $\Pi_t$ which supports $S_t$ at $x_{t+1}$.

To find $x_{t+1}$, take a direction $h \neq 0$ parallel to $\mathrm{Aff}(S_t)$. Since $S_t$ does not contain lines, when moving from $x_t$ either in the direction $h$, or in the direction $-h$, we eventually leave $S_t$, and thus cross the relative boundary of $S_t$. The intersection point is the desired $x_{t+1}$.

3) Set $S_{t+1} = S_t \cap \Pi_t$, replace $t$ with $t+1$ and loop to 1).



Justification: By Important Lemma,

$$\mathrm{Ext}(S_{t+1}) \subset \mathrm{Ext}(S_t),$$

so that

$$\mathrm{Ext}(S_t) \subset \mathrm{Ext}(Q)\ \forall t.$$

Besides this, $\dim(S_{t+1} < \dim(S_t)$, so that Purification algorithm does terminate. the algorithm

Note: Assume you are given a linear form $g^T x$ which is bounded from above on $Q$. Then in the Purification algorithm one can easily ensure that $g^T x_{t+1} \geq g^T x_t$. Thus,

*If $Q$ is a nonempty closed set in $\mathbb{R}^n$ which does not contain lines and $f^T x$ is a linear form which is bounded above on $Q$, then for every point $x_0 \in Q$ there exists (and can be found by Purification) a point $\bar{x} \in \mathrm{Ext}(Q)$ such that $g^T \bar{x} \geq g^T x_0$. In particular, if $g^T x$ attains its maximum on $Q$, then the maximizer can be found among extreme points of $Q$.*

**Proof, $2^0$** If a closed convex set $Q$ contains lines, it has no extreme points.

Another Important Lemma: Let $S$ be a closed convex set such that $\{\bar{x} + th : t \geq 0\} \subset S$ for certain $\bar{x}$. Then

$$\{x + th : t \geq 0\} \subset S \ \forall x \in S.$$

**Proof:** For every $s \geq 0$ and $x \in S$ we have

$$x + sh = \lim_{i \to \infty} \underbrace{[(1 - s/i)x + (s/i)[\bar{x} + th]]}_{\in S}.$$

Note: The set of all directions $h \in \mathbb{R}^n$ such that $\{x + th : t \geq 0\} \subset S$ for some (and then, for all) $x \in S$, is called the *recessive cone* $\mathrm{Rec}(S)$ of closed convex set $S$. $\mathrm{Rec}(S)$ indeed is a cone, and

$$S + \mathrm{Rec}(S) = S.$$

Corollary: If a closed convex set $Q$ contains a line $\ell$, then the parallel lines, passing through points of $Q$, also belong to $Q$. In particular, $Q$ possesses no extreme points.

**Proof, $3^0$:** If a nonempty closed convex set $Q$ is bounded, then $Q = \mathrm{Conv}(\mathrm{Ext}(Q))$.
The inclusion $\mathrm{Conv}(\mathrm{Ext}(Q)) \subset Q$ is evident. Let us prove the opposite inclusion, i.e., prove that every point of $Q$ is a convex combination of extreme points of $Q$.
Induction in $k = \dim Q$. Base $k = 0$ ($Q$ is a singleton) is evident.
Step $k \mapsto k+1$: Given $(k+1)$-dimensional closed and bounded convex set $Q$ and a point $x \in Q$, we, as in the Purification algorithm, can represent $x$ as a convex combination of two points $x_+$ and $x_-$ *from the relative boundary of* $Q$. Let $\Pi_+$ be a hyperplane which supports $Q$ at $x_+$, and let $Q_+ = \Pi_+ \cap Q$. As we know, $Q_+$ is a closed convex set such that

$$\dim Q_+ < \dim Q,\ \mathrm{Ext}(Q_+) \subset \mathrm{Ext}(Q),\ x_+ \in Q_+.$$

Invoking inductive hypothesis,

$$x_+ \in \mathrm{Conv}(\mathrm{Ext}(Q_+)) \subset \mathrm{Conv}(\mathrm{Ext}(Q)).$$

Similarly, $x_- \in \mathrm{Conv}(\mathrm{Ext}(Q))$. Since $x \in [x_-, x_+]$, we get $x \in \mathrm{Conv}(\mathrm{Ext}(Q))$.

## Structure of Polyhedral Sets

♣ Definition: A *polyhedral* set $Q$ in $\mathbb{R}^n$ is a *nonempty* subset in $\mathbb{R}^n$ which is a solution set of a finite system of nonstrict inequalities:

$$Q \text{ is polyhedral} \Leftrightarrow Q = \{x : Ax \geq b\} \neq \emptyset.$$

♠ *Every polyhedral set is convex and closed.*

Question: When a polyhedral set $Q = \{x : Ax \geq b\}$ contains lines? What are these lines, if any?

Answer: $Q$ contains lines iff $A$ has a nontrivial nullspace:

$$\mathrm{Null}(A) \equiv \{h : Ah = 0\} \neq \{0\}.$$

Indeed, a line $\ell = \{x = \bar{x} + th : t \in \mathbb{R}\}$, $h \neq 0$, belongs to $Q$ iff

$$\begin{array}{rl} & \forall t : A(\bar{x} + th) \geq b \\ \Leftrightarrow & \forall t : tAh \geq b - A\bar{x} \\ \Leftrightarrow & Ah = 0 \ \& \ \bar{x} \in Q. \end{array}$$

Fact: *A polyhedral set $Q = \{x : Ax \geq b\}$ always can be represented as*

$$Q = Q_* + L,$$

*where $Q_*$ is a polyhedral set which does not contain lines and $L$ is a linear subspace. In this representation,*

♢ *$L$ is uniquely defined by $Q$ and coincides with* $\mathrm{Null}(A)$*,*

♢ *$Q_*$ can be chosen, e.g., as*

$$Q_* = Q \cap L^{\perp}$$

Structure of polyhedral set which does not contain lines

♣ Theorem. Let

$$Q = \{x : Ax \geq b\} \neq \emptyset$$

be a polyhedral set which does not contain lines (or, which is the same, $\mathrm{Null}(A) = \{0\}$). Then the set $\mathrm{Ext}(Q)$ of extreme points of $Q$ is *nonempty* and *finite*, and

$$Q = \mathrm{Conv}(\mathrm{Ext}(Q)) + \mathrm{Cone}\,\{r_1, ..., r_S\} \qquad (*)$$

for properly chosen vectors $r_1, ..., r_S$.

Note: $\mathrm{Cone}\,\{r_1, ..., r_s\}$ is exactly the recessive cone of $Q$:

$$\begin{array}{rl} & \mathrm{Cone}\,\{r_1, ..., r_S\} \\ = & \{r : x + tr \in Q \ \forall (x \in Q, t \geq 0)\} \\ = & \{r : Ar \geq 0\}. \end{array}$$

This cone is the trivial cone $\{0\}$ iff $Q$ is a *bounded* polyhedral set (called *polytope*).

♣ Combining the above theorems, we come to the following results:

*A polyhedral set $Q$ always can be represented in the form*

$$Q = \left\{ x = \sum_{i=1}^{I} \lambda_i v_i + \sum_{j=1}^{J} \mu_j w_j : \begin{array}{l} \lambda \geq 0, \mu \geq 0 \\ \sum\limits_i \lambda_i = 1 \end{array} \right\} \quad (!)$$

*where $I$, $J$ are positive integers and $v_1, ..., v_I$, $w_1, ..., w_J$ are appropriately chosen points and directions.*

Vice versa, *every set $Q$ of the form (!) is a polyhedral set.*

Note: Polytopes (bounded polyhedral sets) are *exactly* the sets of form (!) with "trivial $w$-part": $w_1 = ... = w_J = 0$.

$$Q \neq \emptyset, \ \& \ \exists A, b : Q = \{x : Ax \geq b\}$$
$$\Updownarrow$$
$$\exists (I, J, v_1, ..., v_I, w_1, ..., w_J) :$$
$$Q = \left\{ x = \sum_{i=1}^{I} \lambda_i v_i + \sum_{j=1}^{J} \mu_j w_j : \begin{array}{l} \lambda \geq 0, \mu \geq 0 \\ \sum_i \lambda_i = 1 \end{array} \right\}$$

Exercise 1: Is it true that the intersection of two polyhedral sets, if nonempty, is a polyhedral set?

Exercise 2: Is it true that the affine image $\{y = Px + p : x \in Q\}$ of a polyhedral set $Q$ is a polyhedral set?

## Applications to Linear Programming

♣ Consider a *feasible* Linear Programming program

$$\min_x c^T x \text{ s.t. } x \in Q = \{x : Ax \geq b\} \qquad \text{(LP)}$$

<u>Observation:</u> We lose nothing when assuming that $\mathrm{Null}(A) = \{0\}$.

Indeed, we have

$$Q = Q_* + \mathrm{Null}(A),$$

where $Q_*$ is a polyhedral set not containing lines. If $c$ is not orthogonal to $\mathrm{Null}(A)$, then (LP) clearly is unbounded. If $c$ is orthogonal to $\mathrm{Null}(A)$, then (LP) is equivalent to the LP program

$$\min_x c^T x \text{ s.t. } x \in Q_*,$$

and now the matrix in a representation $Q_* = \{x : \widetilde{A}x \geq \widetilde{b}\}$ has trivial nullspace.

<u>Assuming $\mathrm{Null}(A) = \{0\}$</u>, let (LP) be bounded (and thus solvable). Since $Q$ is convex, closed and does not contain lines, among the (nonempty!) set of minimizers the objective on $Q$ there is an extreme point of $Q$.

$$\min_x c^T x \text{ s.t. } x \in Q = \{x : Ax \geq b\} \qquad \text{(LP)}$$

We have proved

<u>Proposition.</u> Assume that (LP) is feasible and bounded (and thus is solvable) and that Null($A$) $= \{0\}$. Then among optimal solutions to (LP) there exists at least one which is an extreme point of $Q$.

<u>Question:</u> How to characterize extreme points of the set

$$Q = \{x : Ax \geq b\} \neq \emptyset$$

provided that $A$ is $m \times n$ matrix with Null($A$) $=$ $\{0\}$?

<u>Answer:</u> Extreme points $\bar{x}$ of $Q$ are fully characterized by the following two properties:

♢ $A\bar{x} \geq b$

♢ Among constraints $Ax \geq b$ which are *active* at $\bar{x}$ (i.e., are satisfied as equalities), there are $n$ linearly independent.

**Justification of the answer, $\Rightarrow$:** If $\bar{x}$ is an extreme point of $Q$, then among the constraints $Ax \geq b$ active at $\bar{x}$ there are $n$ linearly independent.
W.l.o.g., assume that the constraints active at $\bar{x}$ are the first $k$ constraints

$$a_i^T x \geq b_i,\ i = 1, ..., k.$$

We should prove that among $n$-dimensional vectors $a_1, ..., a_k$, there are $n$ linearly independent. Assuming otherwise, there exists a nonzero vector $h$ such that $a_i^T h = 0$, $i = 1, ..., k$, that is,

$$a_i^T [\bar{x} \pm \epsilon h] = a_i^T \bar{x} = b_i,\ i = 1, ..., k$$

for all $\epsilon > 0$. Since the remaining constraints $a_i^T x \geq b_i$, $i > k$, are strictly satisfied at $\bar{x}$, we conclude that

$$a_i^T [\bar{x} \pm \epsilon h] \geq b_i,\ i = k + 1, ..., m$$

for all small enough values of $\epsilon > 0$.
We conclude that $\bar{x} \pm \epsilon h \in Q = \{x : Ax \geq b\}$ for all small enough $\epsilon > 0$. Since $h \neq 0$ and $\bar{x}$ is an extreme point of $Q$, we get a contradiction.

**Justification of the answer, $\Leftarrow$:** If $\bar{x} \in Q$ makes equalities $n$ of the constraints $a_i^T x \geq b_i$ with linearly independent vectors of coefficients, then $\bar{x} \in \text{Ext}(Q)$.

W.l.o.g., assume that $n$ active at $\bar{x}$ constraints with linearly independent vectors of coefficients are the first $n$ constraints

$$a_i^T x \geq b_i,\ i = 1, ..., n.$$

We should prove that if $h$ is such that $\bar{x} \pm h \in Q$, then $h = 0$. Indeed, we have

$$\bar{x} \pm h \in Q \Rightarrow a_i^T[\bar{x} \pm h] \geq b_i,\ i = 1, ..., n;$$

since $a_i^T \bar{x} = b_i$ for $i \leq n$, we get

$$a_i^T \bar{x} \pm a_i^T h = a_i^T[\bar{x} \pm h] \geq a_i^T \bar{x},\ i = 1, ..., n,$$

whence

$$a_i^T h = 0,\ i = 1, ..., n. \qquad (*)$$

Since $n$-dimensional vectors $a_1, ..., a_n$ are linearly independent, $(*)$ implies that $h = 0$, Q.E.D.

## Convex Functions

Definition: Let $f$ be a real-valued function defined on a nonempty subset Dom$f$ in $\mathbb{R}^n$. $f$ is called *convex*, if

$\diamondsuit$Dom$f$ is a convex set

$\diamondsuit$for all $x, y \in$ Dom$f$ and $\lambda \in [0, 1]$ one has

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

Equivalent definition: Let $f$ be a real-valued function defined on a nonempty subset Dom$f$ in $\mathbb{R}^n$. The function is called convex, if its *epigraph* – the set

$$\text{Epi}\{f\} = \{(x, t) \in \mathbb{R}^{n+1} : f(x) \leq t\}$$

is a convex set in $\mathbb{R}^{n+1}$.

What does the definition of convexity actually mean?

The inequality

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad (*)$$

where $x, y \in \mathrm{Dom} f$ and $\lambda \in [0,1]$ is automatically satisfied when $x = y$ or when $\lambda = 0/1$. Thus, it says something only when the points $x, y$ are distinct from each other and the point $z = \lambda x + (1-\lambda)y$ is a (relative) interior point of the segment $[x, y]$. What does $(*)$ say in this case?

$\diamondsuit$Observe that $z = \lambda x + (1-\lambda)y = x + (1-\lambda)(y-x)$, whence

$$\|y-x\| : \|y-z\| : \|z-x\| = 1 : \lambda : (1-\lambda)$$

Therefore

$$f(z) \leq \lambda f(x) + (1-\lambda)f(y) \qquad (*)$$

$$\Updownarrow$$

$$f(z) - f(x) \leq \underbrace{(1-\lambda)}_{\frac{\|z-x\|}{\|y-x\|}}(f(y) - f(x))$$

$$\Updownarrow$$

$$\frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(x)}{\|y-x\|}$$

Similarly,

$$f(z) \leq \lambda f(x) + (1-\lambda) f(y) \qquad (*)$$

$$\Updownarrow$$

$$\underbrace{\lambda}_{\frac{\|y-z\|}{\|y-x\|}} (f(y) - f(x)) \leq f(y) - f(z)$$

$$\Updownarrow$$

$$\frac{f(y)-f(x)}{\|y-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}$$

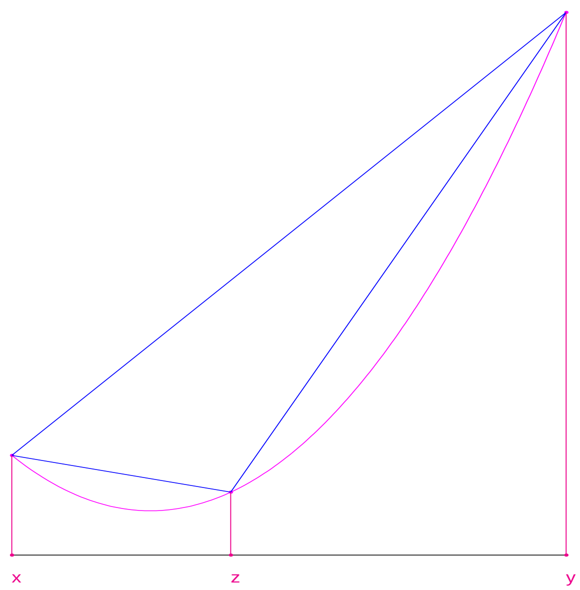
Conclusion: $f$ is convex iff for every three distinct points $x, y, z$ such that $x, y \in \mathrm{Dom}f$ and $z \in [x, y]$, we have $z \in \mathrm{Dom}f$ and

$$\frac{f(z) - f(x)}{\|z - x\|} \leq \frac{f(y) - f(x)}{\|y - x\|} \leq \frac{f(y) - f(z)}{\|y - z\|} \quad (*)$$

Note: From 3 inequalities in $(*)$:

$$\frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(x)}{\|y-x\|}$$
$$\frac{f(y)-f(x)}{\|y-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}$$
$$\frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}$$

every single one implies the other two.



Jensen's Inequality: Let $f(x)$ be a convex function. Then

$$x_i \in \mathrm{Dom} f, \lambda_i \geq 0, \sum_i \lambda_i = 1 \Rightarrow$$
$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$$

Proof: The points $(x_i, f(x_i))$ belong to $\mathrm{Epi}\{f\}$. Since this set is convex, the point

$$(\sum_i \lambda_i x_i, \sum_i \lambda_i f(x_i)) \in \mathrm{Epi}\{f\}.$$

By definition of the epigraph, it follows that

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i).$$

Extension: Let $f$ be convex, $\mathrm{Dom} f$ be closed and $f$ be continuous on $\mathrm{Dom} f$. Consider a probability distribution $\pi(dx)$ supported on $\mathrm{Dom} f$. Then

$$f(\mathbf{E}_\pi\{x\}) \leq \mathbf{E}_\pi\{f(x)\}.$$

Examples:

$\diamondsuit$Functions convex on $\mathbb{R}$: $\bullet$ $x^2$, $x^4$, $x^6$,...

$\bullet$ $\exp\{x\}$

Nonconvex functions on $\mathbb{R}$: $\bullet$ $x^3$ $\bullet$ $\sin(x)$

$\diamondsuit$Functions convex on $\mathbb{R}_+$: $\bullet$ $x^p$, $p \geq 1$

$\bullet$ $-x^p$, $0 \leq p \leq 1$ $\bullet$ $x \ln x$

$\diamondsuit$Functions convex on $\mathbb{R}^n$: $\bullet$ affine function $f(x) = f^T c$

$\diamondsuit$A norm $\|\cdot\|$ on $\mathbb{R}^n$ is a convex function:

$$\begin{array}{rcl}\|\lambda x + (1-\lambda)y\| & \leq & \|\lambda x\| + \|(1-\lambda)y\| \\ & & \text{[Triangle inequality]} \\ & = & \lambda\|x\| + (1-\lambda)\|y\| \\ & & \text{[homogeneity]}\end{array}$$

Application of Jensen's Inequality: Let $p = \{p_i > 0\}_{i=1}^n$, $q = \{q_i > 0\}_{i=1}^n$ be two discrete probability distributions. Claim: The *Kullback-Liebler distance*

$$\sum_i p_i \ln \frac{p_i}{q_i}$$

between the distributions is $\geq 0$.

Indeed, the function $f(x) = -\ln x$, $\mathrm{Dom} f = \{x > 0\}$, is convex. Setting $x_i = q_i/p_i$, $\lambda_i = p_i$ we have

$$\begin{array}{rcl} 0 & = & -\ln\left(\sum\limits_i q_i\right) = f(\sum\limits_i p_i x_i) \\ & \leq & \sum\limits_i p_i f(x_i) = \sum\limits_i p_i(-\ln q_i/p_i) \\ & = & \sum\limits_i p_i \ln(p_i/q_i) \end{array}$$

What is the value of a convex function outside its domain?

<u>Convention.</u> To save words, it is convenient to think that a convex function $f$ is defined *everywhere* on $\mathbb{R}^n$ and takes real values *and value* $+\infty$. With this interpretation, $f$ "remembers" its domain:

$$\begin{array}{c}\mathrm{Dom} f = \{x : f(x) \in \mathbb{R}\} \\ x \notin \mathrm{Dom} f \Rightarrow f(x) = +\infty\end{array}$$

and the definition of convexity becomes

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \ \forall \begin{array}{c} x, y \in \mathbb{R}^n \\ \lambda \in [0,1] \end{array}$$

where the arithmetics of $+\infty$ and reals is given by the rules

$$\begin{array}{l} a \in \mathbb{R} \Rightarrow a + (+\infty) = (+\infty) + (+\infty) = +\infty \\ 0 \cdot (+\infty) = +\infty \\ \lambda > 0 \Rightarrow \lambda \cdot (+\infty) = +\infty \end{array}$$

<u>Note:</u> Operations like $(+\infty) - (+\infty)$ or $(-5) \cdot (+\infty)$ are undefined!

♣ Convexity-preserving operations:

♢Taking conic combinations: If $f_i(x)$ are convex function on $\mathbb{R}^n$ and $\lambda_i \geq 0$, then the function $\sum_i \lambda_i f_i(x)$ is convex

♢Affine substitution of argument: If $f(x)$ is convex function on $\mathbb{R}^n$ and $x = Ay + b$ is an affine mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$, then the function $g(y) = f(Ax + b)$ is convex on $\mathbb{R}^m$

♢Taking supremum: If $f_\alpha(x)$, $\alpha \in \mathcal{A}$, is a family of convex function on $\mathbb{R}^n$, then the function $\sup\limits_{\alpha \in \mathcal{A}} f_\alpha(x)$ is convex.

Proof: $\mathrm{Epi}\{\sup\limits_\alpha f_\alpha(\cdot)\} = \bigcap\limits_\alpha \mathrm{Epi}\{f_\alpha(\cdot)\}$, and intersections of convex sets are convex.

♢Superposition Theorem: Let $f_i(x)$ be convex functions on $\mathbb{R}^n$, $i = 1, ..., m$, and $F(y_1, ..., y_m)$ be a convex and *monotone* function on $\mathbb{R}^m$. Then the function

$$g(x) = \begin{cases} F(f_1(x), ..., f_m(x)) & , x \in \mathrm{Dom}\, f_i, \ \forall i \\ +\infty & , \text{otherwise} \end{cases}$$

is convex.

$\diamondsuit$Partial minimization: Let $f(x,y)$ be a convex function of $z=(x,y)\in\mathbb{R}^n$, and let

$$g(x)=\inf_y f(x,y)$$

be $>-\infty$ for all $x$. Then the function $g(x)$ is convex.

**Proof:** $g$ clearly takes real values and value $+\infty$. Let us check the Convexity Inequality

$$g(\lambda x'+(1-\lambda)x'')\leq\lambda g(x')+(1-\lambda)g(x'') \qquad [\lambda\in[0,1]]$$

There is nothing to check when $\lambda=0$ or $\lambda=1$, so let $0<\lambda<1$. In this case, there is nothing to check when $g(x')$ or $g(x'')$ is $+\infty$, so let $g(x')<+\infty$, $g(x'')<+\infty$. Since $g(x')<+\infty$, for every $\epsilon>0$ there exists $y'$ such that $f(x',y')\leq g(x')+\epsilon$. Similarly, there exists $y''$ such that $f(x'',y'')\leq g(x'')+\epsilon$. Now,

$$\begin{array}{ll} & g(\lambda x'+(1-\lambda)x'')\\ \leq & f(\lambda x'+(1-\lambda)x'',\lambda y'+(1-\lambda)y'')\\ \leq & \lambda f(x',y')+(1-\lambda)f(x'',y'')\\ \leq & \lambda(g(x')+\epsilon)+(1-\lambda)(g(x'')+\epsilon)\\ = & \lambda g(x')+(1-\lambda)g(x'')+\epsilon \end{array}$$

Since $\epsilon>0$ is arbitrary, we get

$$g(\lambda x'+(1-\lambda)x'')\leq\lambda g(x')+(1-\lambda)g(x'').$$

## How to detect convexity?

<u>Convexity is one-dimensional property:</u> A set $X \subset \mathbb{R}^n$ is convex iff the set

$$\{t : a + th \in X\}$$

is, for every $(a, h)$, a convex set on the axis

A function $f$ on $\mathbb{R}^n$ is convex iff the function

$$\phi(t) = f(a + th)$$

is, for every $(a, h)$, a convex function on the axis.

♣ When a function $\phi$ on the axis is convex? Let $\phi$ be convex and finite on $(a,b)$. This is exactly the same as

$$\frac{\phi(z)-\phi(x)}{z-x} \le \frac{\phi(y)-\phi(x)}{y-x} \le \frac{\phi(y)-\phi(z)}{y-z}$$

when $a < x < z < y < b$. Assuming that $\phi'(x)$ and $\phi'(y)$ exist and passing to limits as $z \to x+0$ and $z \to y-0$, we get

$$\phi'(x) \le \frac{\phi(y)-\phi(x)}{y-x} \le \phi'(y)$$

that is, $\phi'(x)$ *is nondecreasing on the set of points from* $(a,b)$ *where it exists*.

The following conditions are necessary and sufficient for convexity of a univariate function:

$\diamondsuit$The domain of the function $\phi$ should be an open interval $\Delta = (a, b)$, possibly with added endpoint(s) (provided that the corresponding endpoint(s) is/are finite)

$\diamondsuit$$\phi$ should be continuous on $(a, b)$ and differentiable everywhere, except, perhaps, a countable set, *and the derivative should be monotonically non-decreasing*

$\diamondsuit$at endpoint of $(a, b)$ which belongs to Dom$\phi$, $\phi$ is allowed to "jump up", but not to jump down.

♣ Sufficient condition for convexity of a univariate function $\phi$: Dom$\phi$ is convex, $\phi$ is continuous on Dom$\phi$ and is twice differentiable, *with nonnegative* $\phi''$, on intDom$\phi$.

Indeed, we should prove that under the condition, if $x < z < y$ are in Dom$\phi$, then

$$\frac{\phi(z)-\phi(x)}{z-x} \leq \frac{\phi(y)-\phi(z)}{y-z}$$

By Lagrange Theorem, the left ratio is $\phi'(\xi)$ for certain $\xi \in (x,z)$, and the right ratio is $\phi'(\eta)$ for certain $\eta \in (z,y)$. Since $\phi''(\cdot) \geq 0$ and $\eta > \xi$, we have $\phi'(\eta) \geq \phi'(\xi)$, Q.E.D.

♣ Sufficient condition for convexity of a multivariate function $f$: Dom$f$ is convex, $f$ is continuous on Dom$f$ and is twice differentiable, *with positive semidefinite Hessian matrix* $f''$, on intDom$f$.

Instructive example: The function $f(x) = \ln(\sum\limits_{i=1}^{n} \exp\{x_i\})$ is convex on $\mathbb{R}^n$.

Indeed,

$$
\begin{array}{rcl}
h^T f'(x) & = & \dfrac{\sum\limits_i \exp\{x_i\} h_i}{\sum\limits_i \exp\{x_i\}} \\
h^T f''(x) h & = & -\dfrac{\left(\sum\limits_i \exp\{x_i\} h_i\right)^2}{\left(\sum\limits_i \exp\{x_i\}\right)^2} + \dfrac{\sum\limits_i \exp\{x_i\} h_i^2}{\sum\limits_i \exp\{x_i\}}
\end{array}
$$

$$h^T f''(x)h = -\left(\frac{\sum\limits_i \text{exp}\{x_i\}h_i}{\sum\limits_i \text{exp}\{x_i\}}\right)^2 + \frac{\sum\limits_i \text{exp}\{x_i\}h_i^2}{\sum\limits_i \text{exp}\{x_i\}}$$

Setting $p_i = \frac{\text{exp}\{x_i\}}{\sum\limits_j \text{exp}\{x_j\}}$, we have

$$\begin{aligned}
h^T f''(x)h &= \sum_i p_i h_i^2 - \left(\sum_i p_i h_i\right)^2 \\
&= \sum_i p_i h_i^2 - \left(\sum_i \sqrt{p_i}(\sqrt{p_i}h_i)\right)^2 \\
&\geq \sum_i p_i h_i^2 - \left(\sum_i (\sqrt{p_i})^2\right)\left(\sum_i (\sqrt{p_i}h_i)^2\right) \\
&= \sum_i p_i h_i^2 - \left(\sum_i p_i h_i^2\right) = 0
\end{aligned}$$

(note that $\sum\limits_i p_i = 1$)

<u>Corollary:</u> When $c_i > 0$, the function

$$g(y) = \ln\left(\sum_i c_i \exp\{a_i^T y\}\right)$$

is convex.

Indeed,

$$g(y) = \ln\left(\sum_i \exp\{\ln c_i + a_i^T y\}\right)$$

is obtained from the convex function

$$\ln\left(\sum_i \exp\{x_i\}\right)$$

by affine substitution of argument.

Gradient Inequality

Proposition: Let $f$ be a function, $x$ be an interior point of the domain of $f$ and $Q$, $x \in Q$, be a convex set such that $f$ is convex on $Q$. Assume that $f$ is differentiable at $x$. Then

$$\forall y \in Q : f(y) \geq f(x) + (y-x)^T f'(x). \qquad (*)$$

**Proof.** Let $y \in Q$. There is nothing to prove when $y = x$ or $f(y) = +\infty$, thus, assume that $f(y) < \infty$ and $y \neq x$. Let is set $z_\epsilon = x + \epsilon(y-x)$, $0 < \epsilon < 1$. Then $z_\epsilon$ is an interior point of the segment $[x, y]$. Since $f$ is convex, we have

$$\frac{f(y) - f(x)}{\|y-x\|} \geq \frac{f(z_\epsilon) - f(x)}{\|z_\epsilon - x\|} = \frac{f(x + \epsilon(y-x)) - f(x)}{\epsilon\|y-x\|}$$

Passing to limit as $\epsilon \to +0$, we arrive at

$$\frac{f(y) - f(x)}{\|y-x\|} \geq \frac{(y-x)^T f'(x)}{\|y-x\|},$$

as required by $(*)$.

## Lipschitz continuity of a convex function

Proposition: Let $f$ be a convex function, and let $K$ be a *closed* and *bounded* set belonging to relative interior of the domain of $f$. Then $f$ is Lipschitz continuous on $K$, that is, there exists a constant $L < \infty$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \forall x, y \in K.$$

Note: All three assumptions on $K$ are essential, as is shown by the following examples:

$\diamondsuit f(x) = -\sqrt{x}$, $\mathrm{Dom} f = \{x \geq 0\}$, $K = [0, 1]$. Here $K \subset \mathrm{Dom} f$ is closed and bounded, but is not contained in the relative interior of $\mathrm{Dom} f$, and $f$ is *not* Lipschitz continuous on $K$

$\diamondsuit f(x) = x^2$, $\mathrm{Dom} f = K = \mathbb{R}$. Here $K$ is closed and belongs to $\mathrm{rint}\,\mathrm{Dom} f$, but is unbounded, and $f$ is *not* Lipschitz continuous on $K$

$\diamondsuit f(x) = \frac{1}{x}$, $\mathrm{Dom} f = \{x > 0\}$, $K = (0, 1]$. Here $K$ is bounded and belongs to rint $Dom f$, but is not closed, and $f$ is *not* Lipschitz continuous on $K$

## Maxima and Minima of Convex Functions

(!) Proposition ["unimodality"] Let $f$ be a convex function and $x_*$ be a local minimizer of $f$:

$$x_* \in \mathrm{Dom} f \\ \& \\ \exists r > 0 : f(x) \geq f(x_*) \; \forall (x : \|x - x_*\| \leq r).$$

Then $x_*$ is a global minimizer of $f$:

$$f(x) \geq f(x_*) \; \forall x.$$

**Proof:** All we need to prove is that if $x \neq x_*$ and $x \in \mathrm{Dom} f$, then $f(x) \geq f(x_*)$. To this end let $z \in (x_*, x)$. By convexity we have

$$\frac{f(z) - f(x_*)}{\|z - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|}.$$

When $z \in (x_*, x)$ is close enough to $x_*$, we have $\frac{f(z)-f(x_*)}{\|z-x_*\|} \geq 0$, whence $\frac{f(x)-f(x_*)}{\|x-x_*\|} \geq 0$, that is, $f(x) \geq f(x_*)$.

Proposition Let $f$ be a convex function. The set of $X_*$ of global minimizers is convex.

**Proof:** This is an immediate corollary of important

Lemma: Let $f$ be a convex function. Then the level sets of $f$, that is, the sets

$$X_a = \{x : f(x) \leq a\}$$

where $a$ is a real, are convex.

**Proof of Lemma:** If $x, y \in X_a$ and $\lambda \in [0,1]$, then

$$\begin{array}{rcl} f(\lambda x + (1-\lambda)y) & \leq & \lambda f(x) + (1-\lambda)f(y) \\ & \leq & \lambda a + (1-\lambda)a = a. \end{array}$$

Thus, $[x, y] \subset X_a$.

♣ *When the minimizer of a convex function is unique?*

Definition: A convex function is called *strictly convex*, if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

whenever $x \neq y$ and $\lambda \in (0,1)$.

Note: *If a convex function $f$ has open domain and is twice continuously differentiable on this domain with*

$$h^T f''(x) h > 0 \quad \forall (x \in \mathrm{Dom} f, h \neq 0),$$

*then $f$ is strictly convex.*

Proposition: For a strictly convex function $f$ a minimizer, if it exists, is unique.

**Proof.** Assume that $X_* = \mathrm{Argmin}\, f$ contains two distinct points $x', x''$. By strong convexity,

$$f(\frac{1}{2}x' + \frac{1}{2}x'') < \frac{1}{2}\left[f(x') + f(x'')\right] = \inf_x f,$$

which is impossible.

Theorem [Optimality conditions in convex minimization] Let $f$ be a function which is differentiable at a point $x_*$ and is convex on a convex set $Q \subset \mathrm{Dom} f$ which contains $x_*$. A necessary and sufficient condition for $f$ to attain its minimum on $Q$ at $x_*$ is

$$(x - x_*)^T f'(x_*) \geq 0 \quad \forall x \in Q. \qquad (*)$$

**Proof, $\Leftarrow$:** Assume that $(*)$ is valid, and let us verify that $f(x) \geq f(x_*)$ for every $x \in Q$. There is nothing to prove when $x = x_*$, thus, let $f(x) < \infty$ and $x \neq x_*$. For $z_\lambda = x_* + \lambda(x - x_*)$ we have

$$\frac{f(z_\lambda) - f(x_*)}{\|z_\lambda - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|} \quad \forall \lambda \in (0,1)$$

or, which is the same,

$$\frac{f(x_* + \lambda[x - x_*]) - f(x_*)}{\lambda\|x - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|} \forall \lambda \in (0,1)$$

As $\lambda \to +0$, the left ratio converges to $(x - x_*)^T f'(x_*)/\|x - x_*\| \geq 0$; thus, $\frac{f(x) - f(x_*)}{\|x - x_*\|} \geq 0$, whence $f(x) \geq f(x_*)$.

"*Let $f$ be a function which is differentiable at a point $x_*$ and is convex on a convex set $Q \subset \mathrm{Dom} f$ which contains $x_*$. A necessary and sufficient condition for $f$ to attain its minimum on $Q$ at $x_*$ is*

$$(x - x_*)^T f'(x_*) \geq 0 \quad \forall x \in Q."$$

**Proof, $\Rightarrow$:** Given that $x_* \in \mathrm{Argmin}_{y \in Q} f(y)$, let $x \in Q$. Then

$$0 \leq \frac{f(x_* + \lambda[x - x_*]) - f(x_*)}{\lambda} \quad \forall \lambda \in (0, 1),$$

whence $(x - x_*)^T f'(x_*) \geq 0$.

♣ Equivalent reformulation: *Let $f$ be a function which is differentiable at a point $x_*$ and is convex on a convex set $Q \subset \mathrm{Dom} f$, $x_* \in Q$. Consider the radial cone of $Q$ at $x_*$:*

$$T_Q(x_*) = \{h : \exists t > 0 : x_* + th \in Q\}$$

*Note: $T_Q(x_*)$ is indeed a cone which is comprised of all vectors of the form $s(x - x_*)$, where $x \in Q$ and $s \geq 0$.*

*$f$ attains its minimum on $Q$ at $x_*$ iff*

$$h^T f'(x_*) \geq 0 \ \forall h \in T_Q(x_*),$$

*or, which is the same, iff*

$$f'(x_*) \in \underbrace{N_Q(x_*) = \{g : g^T h \geq 0 \forall h \in T_Q(x_*)\}}_{\text{normal cone of } Q \text{ at } x_*}. \qquad (*)$$

Example I: $x_* \in \mathrm{int} Q$. Here $T_Q(x_*) = \mathbb{R}^n$, whence $N_Q(x_*) = \{0\}$, and $(*)$ becomes the Fermat equation

$$f'(x_*) = 0$$

Example II: $x_* \in \operatorname{rint} Q$. Let $\mathrm{Aff}(Q) = x_* + L$, where $L$ is a linear subspace in $\mathbb{R}^n$. Here $T_Q(x_*) = L$, whence $N_Q(x_*) = L^\perp$. $(*)$ becomes the condition

$$f'(x_*) \text{ is orthogonal to } L.$$

Equivalently: Let $\mathrm{Aff}(Q) = \{x : Ax = b\}$. Then $L = \{x : Ax = 0\}$, $L^\perp = \{y = A^T\lambda\}$, and the optimality condition becomes

$$\exists \lambda^* : \begin{array}{c} \nabla\Big|_{x=x_*}[f(x) + (\lambda^*)^T(Ax - b)] = 0 \\ \Updownarrow \\ f'(x_*) + \sum\limits_i \lambda_i^* \nabla(a_i^T x - b_i) = 0 \end{array}$$

$$[A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}]$$

Example III: $Q = \{x : Ax - b \le 0\}$ is polyhedral.

In this case,

$$\begin{array}{rl} & T_Q(x_*) \\ = & \left\{h : a_i^T h \le 0 \ \forall i \in I(x_*) = \{i : a_i^T x_* - b_i = 0\}\right\}. \end{array}$$

By Homogeneous Farkas Lemma,

$$\begin{array}{rcl} N_Q(x_*) & \equiv & \{y : a_i^T h \le 0,\ i \in I(x_*) \Rightarrow y^T h \ge 0\} \\ & = & \{y = -\sum\limits_{i \in I(x_*)} \lambda_i a_i : \lambda_i \ge 0\} \end{array}$$

and the optimality condition becomes

$$\exists (\lambda_i^* \ge 0, i \in I(x_*)) : f'(x_*) + \sum_{i \in I(x_*)} \lambda_i^* a_i = 0$$

or, which is the same:

$$\exists \lambda^* \ge 0 : \left\{ \begin{array}{rcl} f'(x_*) + \sum\limits_{i=1}^{m} \lambda_i^* a_i & = & 0 \\ \lambda_i^* (a_i^T x_* - b_i) & = & 0,\ i = 1, ..., m \end{array} \right.$$

The point is that in the *convex* case are necessary *and sufficient* for $x_*$ to be a minimizer of $f$ on $Q$.

Example: Let us solve the problem

$$\min_x \left\{ c^T x + \sum_{i=1}^m x_i \ln x_i : x \geq 0, \sum_i x_i = 1 \right\}.$$

The objective is convex, the domain $Q = \{x \geq 0, \sum_i x_i = 1\}$ is convex (and even polyhedral). *Assuming* that the minimum is achieved at a point $x_* \in \mathrm{rint}\, Q$, the optimality condition becomes

$$\nabla \left[ c^T x + \sum_i x_i \ln x_i + \lambda [\sum_i x_i - 1] \right] = 0$$
$$\Updownarrow$$
$$\ln x_i = -c_i - \lambda - 1 \ \forall i$$
$$\Updownarrow$$
$$x_i = \exp\{1 - \lambda\} \exp\{-c_i\}$$

Since $\sum_i x_i$ should be 1, we arrive at

$$x_i = \frac{\exp\{-c_i\}}{\sum_j \exp\{-c_j\}}.$$

At this point, the optimality condition is satisfied, so that the point indeed is a minimizer.

## Maxima of convex functions

<u>Proposition.</u> Let $f$ be a convex function. Then

$\diamondsuit$If $f$ attains its maximum over $\mathrm{Dom} f$ at a point $x^* \in \mathrm{rint}\,\mathrm{Dom} f$, then $f$ is constant on $\mathrm{Dom} f$

$\diamondsuit$If $\mathrm{Dom} f$ is closed and does not contain lines and $f$ attains its maximum on $\mathrm{Dom} f$, then among the maximizers there is an extreme point of $\mathrm{Dom} f$

$\diamondsuit$If $\mathrm{Dom} f$ is polyhedral and $f$ is bounded from above on $\mathrm{Dom} f$, then $f$ attains its maximum on $\mathrm{Dom} f$.

Subgradients of convex functions

♣ Let $f$ be a convex function and $\bar{x} \in \mathrm{int}\mathrm{Dom} f$. If $f$ differentiable at $\bar{x}$, then, by Gradient Inequality, there exists an affine function, specifically,

$$h(x) = f(\bar{x}) + (x - \bar{x})^T f'(\bar{x}),$$

such that

$$f(x) \geq h(x) \forall x \;\&\; f(\bar{x}) = h(\bar{x}) \qquad (*)$$

Affine function with property $(*)$ may exist also in the case when $f$ is *not* differentiable at $\bar{x} \in \mathrm{Dom} f$. $(*)$ implies that

$$h(x) = f(\bar{x}) + (x - \bar{x})^T g \qquad (**)$$

for certain $g$. Function $(**)$ indeed satisfies $(*)$ if and only if $g$ is such that

$$f(x) \geq f(\bar{x}) + (x - \bar{x})^T g \quad \forall x \qquad (!)$$

Definition. Let $f$ be a convex function and $\bar{x} \in \mathrm{Dom} f$. Every vector $g$ satisfying

$$f(x) \geq f(\bar{x}) + (x - \bar{x})^T g \quad \forall x \qquad (!)$$

is called a *subgradient* of $f$ at $\bar{x}$. The set of all subgradients, if any, of $f$ at $\bar{x}$ is called *subdifferential* $\partial f(\bar{x})$ of $f$ at $\bar{x}$.

Example I: By Gradient Inequality, if convex function $f$ is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) \in \partial f(\bar{x})$. If, in addition, $\bar{x} \in \mathrm{int Dom} f$, then $\nabla f(\bar{x})$ is the *unique* element of $\partial f(\bar{x})$.

Example II: Let $f(x) = |x|$ $(x \in \mathbb{R})$. When $\bar{x} \neq 0$, $f$ is differentiable at $\bar{x}$, whence $\partial f(\bar{x}) = f'(\bar{x})$. When $\bar{x} = 0$, subgradients $g$ are given by

$$|x| \geq 0 + gx = gx \;\forall x,$$

that is, $\partial f(0) = [-1, 1]$.

Note: In the case in question, $f$ has directional derivative

$$Df(x)[h] = \lim_{t \to +0} \frac{f(x + th) - f(x)}{t}$$

at every point $x \in \mathbb{R}$ along every direction $h \in \mathbb{R}$, and this derivative is nothing but

$$Df(x)[h] = \max_{g \in \partial f(x)} g^T h$$

<u>Proposition:</u> Let $f$ be convex. Then

♢For every $x \in \mathrm{Dom} f$, the subdifferential $\partial f(x)$ is closed convex set

♢If $x \in \mathrm{rint}\,\mathrm{Dom} f$, then $\partial f(x)$ is nonempty.

♢If $x \in \mathrm{rint}\,\mathrm{Dom} f$, then, for every $h \in \mathbb{R}^n$,

$$\exists Df(x)[h] \equiv \lim_{t \to +0} \frac{f(x+th) - f(x)}{t} = \max_{g \in \partial f(x)} g^T h.$$

♢Assume that $\bar{x} \in \mathrm{Dom} f$ is represented as $\lim\limits_{i \to \infty} x_i$ with $x_i \in \mathrm{Dom} f$ and that

$$f(\bar{x}) \leq \liminf_{i \to \infty} f(x_i)$$

If a sequence $g_i \in \partial f(x_i)$ converges to certain vector $g$, then $g \in \partial f(\bar{x})$.

♢The multi-valued mapping $x \mapsto \partial f(x)$ is locally bounded at every point $\bar{x} \in \mathrm{int}\mathrm{Dom} f$, that is, whenever $\bar{x} \in \mathrm{int}\mathrm{Dom} f$, there exist $r > 0$ and $R < \infty$ such that

$$\|x - \bar{x}\|_2 \leq r, g \in \partial f(x) \Rightarrow \|g\|_2 \leq R.$$

**Selected proof:** "If $\bar{x} \in \operatorname{rint} \operatorname{Dom} f$, then $\partial f(\bar{x})$ is nonempty."
W.l.o.g. let $\operatorname{Dom} f$ be full-dimensional, so that $\bar{x} \in \operatorname{int} \operatorname{Dom} f$. Consider the convex set

$$T = \text{Epi}\{f\} = \{(x, t) : t \geq f(x)\}.$$

Since $f$ is convex, it is continuous on $\operatorname{int} \operatorname{Dom} f$, whence $T$ has a nonempty interior. The point $(\bar{x}, f(\bar{x}))$ clearly does not belong to this interior, whence $S = \{(\bar{x}, f(\bar{x}))\}$ can be separated from $T$: there exists $(\alpha, \beta) \neq 0$ such that

$$\alpha^T \bar{x} + \beta f(\bar{x}) \leq \alpha^T x + \beta t \quad \forall (x, t \geq f(x)) \quad (*)$$

Clearly $\beta \geq 0$ (otherwise $(*)$ will be impossible when $x = \bar{x}$ and $t > f(\bar{x})$ is large).
Claim: $\beta > 0$. Indeed, with $\beta = 0$, (*) implies

$$\alpha^T \bar{x} \leq \alpha^T x \; \forall x \in \operatorname{Dom} f \qquad (**)$$

Since $(\alpha, \beta) \neq 0$ and $\beta = 0$, we have $\alpha \neq 0$; but then $(**)$ contradicts $\bar{x} \in \operatorname{int} \operatorname{Dom} f$.
$\diamondsuit$Since $\beta > 0$, $(*)$ implies that if $g = \beta^{-1}\alpha$, then

$$g^T \bar{x} + f(\bar{x}) \leq g^T x + f(x) \; \forall x \in \operatorname{Dom} f,$$

that is,

$$f(x) \geq f(\bar{x}) + (x - \bar{x})^T g \; \forall x.$$

## Elementary Calculus of Subgradients

$\diamondsuit$If $g_i \in \partial f_i(x)$ and $\lambda_i \geq 0$, then

$$\sum_i \lambda_i g_i \in \partial(\sum_i \lambda_i f_i)(x)$$

$\diamondsuit$If $g_\alpha \in \partial f_\alpha(x)$, $\alpha \in \mathcal{A}$,

$$f(\cdot) = \sup_{\alpha \in \mathcal{A}} f_\alpha(\cdot)$$

and

$$f(x) = f_\alpha(x),\ \alpha \in \mathcal{A}_*(x) \neq \emptyset,$$

then every convex combination of vectors $g_\alpha$, $\alpha \in \mathcal{A}_*(x)$, is a subgradient of $f$ at $x$

$\diamondsuit$If $g_i \in \text{dom} f_i(x)$, $i = 1, ..., m$, and $F(y_1, ..., y_m)$ is convex and monotone and $0 \leq d \in \partial F(f_1(x), ..., f_m(x))$, then the vector

$$\sum_i d_i g_i$$

is a subgradient of $F(f_1(\cdot), ..., f_m(\cdot))$ at $x$.

Convex Programming
Lagrange Duality
Saddle Points

♣ Mathematical Programming program is

$$f_* = \min_x \left\{ f(x) : \begin{array}{rl} g(x) \equiv (g_1(x), ..., g_m(x))^T & \leq 0 \\ h(x) = (h_1(x), ..., h_k(x))^T & = 0 \\ x & \in X \end{array} \right\} \quad (P)$$

♢$x$ is the *design vector*. Values of $x$ are called *solutions* to $(P)$

♢$f(x)$ is the *objective*

♢$g(x) \equiv (g_1(x), ..., g_m(x))^T \leq 0$ – *inequality constraints*

♢$h(x) = (h_1(x), ..., h_k(x))^T = 0$ – *equality constraints*

♢$X \subset \mathbb{R}^n$ – *domain*. We always assume that the objective and the constraints are well-defined on $X$.

$$f_* = \min_x \left\{ f(x) : \begin{array}{rl} g(x) \equiv (g_1(x), ..., g_m(x))^T & \leq 0 \\ h(x) = (h_1(x), ..., h_k(x))^T & = 0 \\ x & \in X \end{array} \right\} \qquad (P)$$

♣ Solution $x$ is called *feasible*, if it satisfies all the constraints. Problem which has feasible solutions is called *feasible*.

♣ If the objective is (below) bounded on the set of feasible solutions, $(P)$ is called *bounded*.

♣ The *optimal value* $f_*$ is

$$f_* = \begin{cases} \inf_x \{f(x) : \; x \text{ is feasible}\}, & (P) \text{ is feasible} \\ +\infty, & \text{otherwise} \end{cases}$$

$f_*$ is a real for feasible and bounded problem, is $-\infty$ for feasible unbounded problem, and is $+\infty$ for infeasible problem.

♣ *Optimal solution* of $(P)$ is a feasible solution $x_*$ such that $f(x_*) = f_*$. Problem which has optimal solutions is called *solvable*.

$$f_* = \min_x \left\{ f(x) : \begin{array}{rl} g(x) \equiv (g_1(x), ..., g_m(x))^T & \leq 0 \\ h(x) = (h_1(x), ..., h_k(x))^T & = 0 \\ x & \in X \end{array} \right\}$$
$$(P)$$

♣ Problem $(P)$ is called *convex*, if

♢ $X$ is a convex subset of $\mathbb{R}^n$

♢ $f(\cdot)$, $g_1(\cdot)$,...,$g_m(\cdot)$ are *convex real-valued* functions on $X$

♢ There are no equality constraints

[we could allow *linear* equality constraints, but this does not add generality]

Preparing tools for Lagrange Duality:
Convex Theorem on Alternative

♣ Question: *How to certify insolvability of the system*

$$\begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0, \ j = 1, ..., m \\ x & \in & X \end{array} \qquad (I)$$

♣ Answer: *Assume that there exist nonnegative weights $\lambda_j$, $j = 1, ..., m$, such that the inequality*

$$f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) < c$$

*has no solutions in $X$:*

$$\exists \lambda_j \geq 0 : \qquad \inf_{x \in X} [f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)] \geq c.$$

*Then $(I)$ is insolvable.*

♣ Convex Theorem on Alternative: Consider a system of constraints on $x$

$$\begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0,\ j=1,...,m \\ x & \in & X \end{array} \qquad (I)$$

along with system of constraints on $\lambda$:

$$\begin{array}{rcl} \inf\limits_{x\in X}[f(x)+\sum\limits_{j=1}^m \lambda_j g_j(x)] & \geq & c \\ \lambda_j & \geq & 0,\ j=1,...,m \end{array} \qquad (II)$$

♢[Trivial part] If (II) is solvable, then (I) is insolvable

♢[Nontrivial part] If (I) is insolvable and system (I) is convex:

— $X$ is convex set

— $f$, $g_1,...,g_m$ are real-valued convex functions on $X$

and the subsystem

$$\begin{array}{rcl} g_j(x) & < & 0,\ j=1,...,m, \\ x & \in & X \end{array}$$

is solvable [Slater condition], then $(II)$ is solvable.

$$\begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0,\ j = 1, ..., m \\ x & \in & X \end{array} \qquad (I)$$

**Proof of Nontrivial part:** Assume that $(I)$ has no solutions. Consider two sets in $\mathbb{R}^{m+1}$:

$$\overbrace{\left\{ u \in \mathbb{R}^{m+1} : \exists x \in X : \begin{array}{rcl} f(x) & \leq & u_0 \\ g_1(x) & \leq & u_1 \\ & .......... & \\ g_m(x) & \leq & u_m \end{array} \right\}}^{T}$$

$$\underbrace{\left\{ u \in \mathbb{R}^{m+1} : u_0 < c, u_1 \leq 0, ..., u_m \leq 0 \right\}}_{S}$$

<u>Observations:</u> $\diamondsuit S$, $T$ are convex and nonempty
$\diamondsuit S$, $T$ do not intersect (otherwise $(I)$ would have a solution)
<u>Conclusion:</u> $S$ and $T$ can be separated:

$$\exists (a_0, ..., a_m) \neq 0 : \inf_{u \in T} a^T u \geq \sup_{u \in S} a^T u$$

$$\overbrace{\left\{u \in \mathbb{R}^{m+1} : \exists x \in X : \begin{array}{rcl} f(x) & \leq & u_0 \\ g_1(x) & \leq & u_1 \\ & \cdots\cdots\cdots & \\ g_m(x) & \leq & u_m \end{array}\right\}}^{T}$$

$$\underbrace{\left\{u \in \mathbb{R}^{m+1} : u_0 < c, u_1 \leq 0, ..., u_m \leq 0\right\}}_{S}$$

$$\exists (a_0, ..., a_m) \neq 0 :$$

$$\inf_{x \in X} \inf_{\begin{array}{rcl} u_0 & \geq & f(x) \\ u_1 & \geq & g_1(x) \\ & \vdots & \\ u_m & \geq & g_m(x) \end{array}} [a_0 u_0 + a_1 u_1 + ... + a_m u_m]$$

$$\geq \sup_{\begin{array}{rcl} u_0 & < & c \\ u_1 & \leq & 0 \\ & \vdots & \\ u_m & \leq & 0 \end{array}} [a_0 u_0 + a_1 u_1 + ... + a_m u_m]$$

<u>Conclusion:</u> $a \geq 0$, whence

$$\inf_{x \in X} [a_0 f(x) + a_1 g_1(x) + ... + a_m g_m(x)] \geq a_0 c.$$

Summary:

$\exists a \geq 0, a \neq 0$ :
$\inf\limits_{x \in X} [a_0 f(x) + a_1 g_1(x) + \ldots + a_m g_m(x)] \geq a_0 c$

Observation: $a_0 > 0$.

Indeed, otherwise $0 \neq (a_1, \ldots, a_m) \geq 0$ and

$$\inf_{x \in X} [a_1 g_1(x) + \ldots + a_m g_m(x)] \geq 0,$$

while $\exists \bar{x} \in X : g_j(\bar{x}) < 0$ for all $j$.

Conclusion: $a_0 > 0$, whence

$$\inf_{x \in X} \left[ f(x) + \sum_{j=1}^{m} \underbrace{\left[\frac{a_j}{a_0}\right]}_{\lambda_j \geq 0} g_j(x) \right] \geq c.$$

## Lagrange Function

♣ Consider optimization program

$$\mathrm{Opt}(P) = \min\left\{ f(x) : g_j(x) \leq 0,\ j \leq m,\ x \in X \right\}. \tag{$P$}$$

and associate with it *Lagrange function*

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

along with the *Lagrange Dual problem*

$$\mathrm{Opt}(D) = \max_{\lambda \geq 0} \underline{L}(\lambda),\ \underline{L}(\lambda) = \inf_{x \in X} L(x, \lambda) \tag{$D$}$$

♣ Convex Programming Duality Theorem:

♢[Weak Duality] For every $\lambda \geq 0$, $\underline{L}(\lambda) \leq \mathrm{Opt}(P)$. In particular,

$$\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$$

♢[Strong Duality] If $(P)$ is convex and below bounded and satisfies Slater condition, then $(D)$ is solvable, and

$$\mathrm{Opt}(D) = \mathrm{Opt}(P).$$

$$\mathrm{Opt}(P) = \min\{f(x) : g_j(x) \le 0,\ j \le m,\ x \in X\} \quad (P)$$
$$\Downarrow$$
$$L(x,\lambda) = f(x) + \sum_j \lambda_j g_j(x)$$
$$\Downarrow$$
$$\mathrm{Opt}(D) = \max_{\lambda \ge 0} \underbrace{\left[\inf_{x\in X} L(x,\lambda)\right]}_{\underline{L}(\lambda)} \qquad (D)$$

Weak Duality: "$\mathrm{Opt}(D) \le \mathrm{Opt}(P)$": There is nothing to prove when $(P)$ is infeasible, that is, when $\mathrm{Opt}(P) = \infty$. If $x$ is feasible for $(P)$ and $\lambda \ge 0$, then $L(x,\lambda) \le f(x)$, whence

$$\begin{array}{rcl}
\lambda \ge 0 \Rightarrow \underline{L}(\lambda) & \equiv & \inf\limits_{x\in X} L(x,\lambda) \\
& \le & \inf\limits_{x\in X \text{ is feasible}} L(x,\lambda) \\
& \le & \inf\limits_{x\in X \text{ is feasible}} f(x) \\
& = & \mathrm{Opt}(P) \\
\Rightarrow \mathrm{Opt}(D) & = & \sup\limits_{\lambda\ge 0} \underline{L}(\lambda) \le \mathrm{Opt}(D).
\end{array}$$

$$\begin{array}{c}
\mathrm{Opt}(P) = \min\left\{f(x) : g_j(x) \le 0,\ j \le m,\ x \in X\right\} \quad (P) \\
\Downarrow \\
L(x,\lambda) = f(x) + \sum_j \lambda_j g_j(x) \\
\Downarrow \\
\mathrm{Opt}(D) = \max_{\lambda \ge 0} \underbrace{\left[\inf_{x \in X} L(x,\lambda)\right]}_{\underline{L}(\lambda)} \qquad (D)
\end{array}$$

Strong Duality: "If $(P)$ is convex and below bounded and satisfies Slater condition, then $(D)$ is solvable and $\mathrm{Opt}(D) = \mathrm{Opt}(P)$":
The system

$$f(x) < \mathrm{Opt}(P),\ g_j(x) \le 0, j = 1, ..., m,\ x \in X$$

has no solutions, while the system

$$g_j(x) < 0, j = 1, ..., m,\ x \in X$$

has a solution. By CTA,

$$\exists \lambda^* \ge 0 : f(x) + \sum_j \lambda_j^* g_j(x) \ge \mathrm{Opt}(P)\ \forall x \in X,$$

whence

$$\underline{L}(\lambda^*) \ge \mathrm{Opt}(P). \qquad (*)$$

Combined with Weak Duality, $(*)$ says that

$$\mathrm{Opt}(D) = \underline{L}(\lambda^*) = \mathrm{Opt}(P).$$

$$
\begin{array}{c}
\mathrm{Opt}(P) = \min\left\{f(x) : g_j(x) \le 0,\ j \le m,\ x \in X\right\} \quad (P) \\
\Downarrow \\
L(x,\lambda) = f(x) + \sum_j \lambda_j g_j(x) \\
\Downarrow \\
\mathrm{Opt}(D) = \max_{\lambda \ge 0} \underbrace{\left[\inf_{x\in X} L(x,\lambda)\right]}_{\underline{L}(\lambda)} \quad (D)
\end{array}
$$

<u>Note:</u> The Lagrange function "remembers", up to equivalence, both $(P)$ and $(D)$.

Indeed,

$$\mathrm{Opt}(D) = \sup_{\lambda\ge 0} \inf_{x\in X} L(x,\lambda)$$

is given by the Lagrange function. Now consider the function

$$\overline{L}(x) = \sup_{\lambda \ge 0} L(x,\lambda) = \begin{cases} f(x), & g_j(x) \le 0,\ j \le m \\ +\infty, & \text{otherwise} \end{cases}$$

$(P)$ clearly is equivalent to the problem of minimizing $\overline{L}(x)$ over $x \in X$:

$$\mathrm{Opt}(P) = \inf_{x\in X} \sup_{\lambda\ge 0} L(x,\lambda)$$

## Saddle Points

♣ Let $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ be nonempty sets, and let $F(x,\lambda)$ be a real-valued function on $X \times \Lambda$. This function gives rise to two optimization problems

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

Game interpretation: Player I chooses $x \in X$, player II chooses $\lambda \in \Lambda$. With choices of the players $x, \lambda$, player I pays to player II the sum of $F(x,\lambda)$. What should the players do to optimize their wealth?

♢If Player I chooses $x$ first, and Player II knows this choice when choosing $\lambda$, II will maximize his profit, and the loss of I will be $\overline{F}(x)$. To minimize his loss, I should solve $(P)$, thus ensuring himself loss $\mathrm{Opt}(P)$ or less.

♢If Player II chooses $\lambda$ first, and Player I knows this choice when choosing $x$, I will minimize his loss, and the profit of II will be $\underline{F}(\lambda)$. To maximize his profit, II should solve $(D)$, thus ensuring himself profit $\mathrm{Opt}(D)$ or more.

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

<u>Observation:</u> For Player I, second situation seems better, so that it is natural to guess that his anticipated loss in this situation is $\leq$ his anticipated loss in the first situation:

$$\mathrm{Opt}(D) \equiv \sup_{\lambda\in\Lambda} \inf_{x\in X} F(x,\lambda) \leq \inf_{x\in X} \sup_{\lambda\in\Lambda} F(x,\lambda) \equiv \mathrm{Opt}(P).$$

This indeed is true: assuming $\mathrm{Opt}(P) < \infty$ (otherwise the inequality is evident),

$$\forall(\epsilon > 0): \quad \exists x_\epsilon \in X : \sup_{\lambda\in\Lambda} F(x_\epsilon,\lambda) \leq \mathrm{Opt}(P) + \epsilon$$
$$\Rightarrow \forall \lambda \in \Lambda : \underline{F}(\lambda) = \inf_{x\in X} F(x,\lambda) \leq F(x_\epsilon,\lambda) \leq \mathrm{Opt}(P) + \epsilon$$
$$\Rightarrow \mathrm{Opt}(D) \equiv \sup_{\lambda\in\Lambda} \underline{F}(\lambda) \leq \mathrm{Opt}(P) + \epsilon$$
$$\Rightarrow \mathrm{Opt}(D) \leq \mathrm{Opt}(P).$$

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

♣ What should the players do when making their choices simultaneously?

A "good case" when we can answer this question – $F$ has a *saddle point*.

Definition: *We call a point* $(x_*,\lambda_*)\in X\times\Lambda$ *a saddle point of* $F$, *if*

$$F(x,\lambda_*)\geq F(x_*,\lambda_*)\geq F(x_*,\lambda)\ \forall(x\in X,\lambda\in\Lambda).$$

In game terms, a saddle point is an *equilibrium* – no one of the players can improve his wealth, provided the adversary keeps his choice unchanged.

Proposition: $F$ has a saddle point if and only if both $(P)$ and $(D)$ are solvable with equal optimal values. In this case, the saddle points of $F$ are exactly the pairs $(x_*,\lambda_*)$, where $x_*$ is an optimal solution to $(P)$, and $\lambda_*$ is an optimal solution to $(D)$.

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

**Proof, $\Rightarrow$:** Assume that $(x_*,\lambda_*)$ is a saddle point of $F$, and let us prove that $x_*$ solves $(P)$, $\lambda_*$ solves $(D)$, and $\mathrm{Opt}(P)=\mathrm{Opt}(D)$. Indeed, we have

$$F(x,\lambda_*) \geq F(x_*,\lambda_*) \geq F(x_*,\lambda)\ \forall(x\in X,\lambda\in\Lambda)$$

whence

$$\begin{array}{l} \mathrm{Opt}(P) \leq \overline{F}(x_*) = \sup\limits_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*,\lambda_*) \\ \mathrm{Opt}(D) \geq \underline{F}(\lambda_*) = \inf\limits_{x\in X} F(x,\lambda_*) = F(x_*,\lambda_*) \end{array}$$

Since $\mathrm{Opt}(P) \geq \mathrm{Opt}(D)$, we see that all inequalities in the chain

$$\mathrm{Opt}(P) \leq \overline{F}(x_*) = F(x_*,\lambda_*) = \underline{F}(\lambda_*) \leq \mathrm{Opt}(D)$$

are equalities. Thus, $x_*$ solves $(P)$, $\lambda_*$ solves $(D)$ and $\mathrm{Opt}(P)=\mathrm{Opt}(D)$.

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

**Proof, $\Leftarrow$.** Assume that $(P)$, $(D)$ have optimal solutions $x_*, \lambda_*$ and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, and let us prove that $(x_*, \lambda_*)$ is a saddle point. We have

$$\begin{array}{l} \mathrm{Opt}(P) = \overline{F}(x_*) = \sup_{\lambda\in\Lambda} F(x_*,\lambda) \geq F(x_*,\lambda_*) \\ \mathrm{Opt}(D) = \underline{F}(\lambda_*) = \inf_{x\in X} F(x,\lambda_*) \leq F(x_*,\lambda_*) \end{array} \quad (*)$$

Since $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, all inequalities in $(*)$ are equalities, so that

$$\sup_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*\lambda_*) = \inf_{x\in X} F(x,\lambda_*).$$

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \leq 0, j \leq m, x \in X \right\} \quad (P)$$
$$\Downarrow$$
$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

Theorem [Saddle Point form of Optimality Conditions in Convex Programming]
Let $x_* \in X$.
$\diamondsuit$[Sufficient optimality condition] If $x_*$ can be extended, by a $\lambda^* \geq 0$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\}$:

$$L(x, \lambda^*) \geq L(x_*, \lambda^*) \geq L(x_*, \lambda) \ \forall (x \in X, \lambda \geq 0),$$

then $x_*$ is optimal for $(P)$.
$\diamondsuit$[Necessary optimality condition] If $x_*$ is optimal for $(P)$ and $(P)$ is convex and satisfies the Slater condition, then $x_*$ can be extended, by a $\lambda^* \geq 0$, to a saddle point of the Lagrange function on $X \times \{\lambda \geq 0\}$.

$$\text{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \le 0, j \le m, x \in X \right\} \quad (P)$$

$$\Downarrow$$

$$L(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

**Proof, $\Rightarrow$:** "Assume $x_* \in X$ and $\exists \lambda^* \ge 0$ :

$L(x,\lambda^*) \ge L(x_*,\lambda^*) \ge L(x_*,\lambda)\ \forall (x \in X, \lambda \ge 0)$.

Then $x_*$ is optimal for $(P)$."

Clearly, $\sup\limits_{\lambda \ge 0} L(x_*,\lambda) = \begin{cases} +\infty, & x_* \text{ is infeasible} \\ f(x_*), & \text{otherwise} \end{cases}$

Thus, $\lambda^* \ge 0$ & $L(x_*,\lambda^*) \ge L(x_*,\lambda)\ \forall \lambda \ge 0$ is equivalent to

$$g_j(x_*) \le 0 \forall j \ \&\ \lambda_j^* g_j(x_*) = 0 \forall j.$$

Consequently, $L(x_*,\lambda^*) = f(x_*)$, whence

$$L(x,\lambda^*) \ge L(x_*,\lambda^*)\ \forall x \in X$$

reads as

$$L(x,\lambda^*) \ge f(x_*)\ \forall x. \qquad (*)$$

Since for $\lambda \ge 0$ one has $f(x) \ge L(x,\lambda)$ for all feasible $x$, $(*)$ implies that

$$x \text{ is feasible } \Rightarrow f(x) \ge f(x_*).$$

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \le 0, j \le m, x \in X \right\} \quad (P)$$

$$\Downarrow$$

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

**Proof, $\Leftarrow$:** Assume $x_*$ is optimal for convex problem $(P)$ satisfying the Slater condition. Then $\exists \lambda^* \ge 0$ :

$$L(x, \lambda^*) \ge L(x_*, \lambda^*) \ge L(x_*, \lambda) \ \forall (x \in X, \lambda \ge 0).$$

By Lagrange Duality Theorem, $\exists \lambda^* \ge 0$:

$$f(x_*) = \underline{L}(\lambda^*) \equiv \inf_{x \in X} \left[ f(x) + \sum_j \lambda_j^* g_j(x) \right]. \qquad (*)$$

Since $x_*$ is feasible, we have

$$\inf_{x \in X} \left[ f(x) + \sum_j \lambda_j^* g_j(x) \right] \le f(x_*) + \sum_j \lambda_j^* g_j(x_*) \le f(x_*).$$

By $(*)$, the last " $\ge$ " here is " $=$ ", which with $\lambda^* \ge 0$ is possible iff $\lambda_j^* g_j(x_*) = 0 \, \forall j$

$$\Rightarrow f(x_*) = L(x_*, \lambda^*) \ge L(x_*, \lambda) \ \forall \lambda \ge 0.$$

Now $(*)$ reads $L(x, \lambda^*) \ge f(x_*) = L(x_*, \lambda^*)$.

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \leq 0, j \leq m, x \in X \right\} \quad (P)$$

$$\Downarrow$$

$$L(x, \lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x)$$

<u>Theorem</u> [Karush-Kuhn-Tucker Optimality Conditions in Convex Programming] Let $(P)$ be a convex program, let $x^*$ be its feasible solution, and let the functions $f$, $g_1$,...,$g_m$ be differentiable at $x^*$. Then

♢The Karush-Kuhn-Tucker condition:

Exist Lagrange multipliers $\lambda^* \geq 0$ such that

$$\begin{array}{l} \nabla f(x_*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x_*) \in N_X^*(x_*) \\ \lambda_j^* g_j(x_*) = 0,\ j \leq m \ \ \text{[complementary slackness]} \end{array}$$

is <u>*sufficient*</u> for $x_*$ to be optimal.

♢If $(P)$ satisfies *restricted Slater condition:*

$\exists \bar{x} \in \mathrm{rint}\, X : g_j(\bar{x}) \leq 0$ for all constraints and $g_j(\bar{x}) < 0$ for all <u>*nonlinear*</u> constraints,

then the KKT is <u>*necessary and sufficient*</u> for $x_*$ to be optimal.

$$\text{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \le 0, j \le m, x \in X \right\} \quad (P)$$

$$\Downarrow$$

$$L(x, \lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x)$$

**Proof, $\Rightarrow$:** Let $(P)$ be convex, $x_*$ be feasible, and $f$, $g_j$ be differentiable at $x_*$. Assume also that the KKT holds:
Exist Lagrange multipliers $\lambda^* \ge 0$ such that

$$(a) \quad \nabla f(x_*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x_*) \in N_X^*(x_*)$$

$$(b) \quad \lambda_j^* g_j(x_*) = 0, \ j \le m \quad \text{[complementary slackness]}$$

Then $x_*$ is optimal.
Indeed, complementary slackness plus $\lambda^* \ge 0$ ensure that

$$L(x_*, \lambda^*) \ge L(x_*, \lambda) \quad \forall \lambda \ge 0.$$

Further, $L(x, \lambda^*)$ is convex in $x \in X$ and differentiable at $x_* \in X$, so that $(a)$ implies that

$$L(x, \lambda^*) \ge L(x_*, \lambda^*) \quad \forall x \in X.$$

Thus, $x_*$ can be extended to a saddle point of the Lagrange function and therefore is optimal for $(P)$.

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : g_j(x) \le 0, j \le m, x \in X \right\} \quad (P)$$
$$\Downarrow$$
$$L(x,\lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x)$$

**Proof, $\Leftarrow$ [under Slater condition]** Let $(P)$ be convex and satisfy the Slater condition, let $x_*$ be optimal and $f$, $g_j$ be differentiable at $x_*$. Then
Exist Lagrange multipliers $\lambda^* \ge 0$ such that

$$(a) \quad \nabla f(x_*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x_*) \in N_X^*(x_*)$$
$$(b) \quad \lambda_j^* g_j(x_*) = 0,\ j \le m \ \text{[complementary slackness]}$$

By Saddle Point Optimality condition, from optimality of $x_*$ it follows that $\exists \lambda^* \ge 0$ such that $(x_*, \lambda^*)$ is a saddle point of $L(x,\lambda)$ on $X \times \{\lambda \ge 0\}$. This is equivalent to

$$\lambda_j^* g_j(x_*) = 0 \ \forall j \ \& \ \underbrace{\min_{x\in X} L(x,\lambda^*) = L(x_*,\lambda^*)}_{(*)}$$

Since the function $L(x,\lambda^*)$ is convex in $x \in X$ and differentiable at $x_* \in X$, relation $(*)$ implies $(a)$.

♣ Application example: Assuming $a_i > 0$, $p \geq 1$, let us solve the problem

$$\min_x \left\{ \sum_i \frac{a_i}{x_i} : x > 0, \sum_i x_i^p \leq 1 \right\}$$

Assuming $x_* > 0$ is a solution such that $\sum_i (x_i^*)^p = 1$, the KKT conditions read

$$\begin{array}{rcl} \nabla_x \left\{ \sum_i \frac{a_i}{x_i} + \lambda(\sum_i x_i^2 - 1) \right\} & = & 0 \Leftrightarrow \frac{a_i}{x_i^2} = p\lambda x_i^{p-1} \\ \sum_i x_i^p & = & 1 \end{array}$$

whence $x_i = c(\lambda) a_i^{\frac{1}{p+1}}$. Since $\sum_i x_i^p$ should be 1, we get

$$x_i^* = \frac{a_i^{\frac{1}{p+1}}}{\left(\sum_j a_j^{\frac{p}{p+1}}\right)^{\frac{1}{p}}}.$$

This point is feasible, problem is convex, KKT at the point is satisfied
$\Rightarrow x^*$ is optimal!

# Existence of Saddle Points

♣ <u>Theorem</u> [Sion-Kakutani] Let $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ be nonempty convex closed sets and $F(x, \lambda) : X \times \Lambda \to \mathbb{R}$ be a continuous function which is convex in $x \in X$ and concave in $\lambda \in \Lambda$.
Assume that $X$ is compact, and that there exists $\bar{x} \in X$ such that all the sets

$$\Lambda_a : \{\lambda \in \Lambda : F(\bar{x}, \lambda) \geq a\}$$

are bounded (e.g., $\Lambda$ is bounded).
Then $F$ possesses a saddle point on $X \to \Lambda$.
<u>Proof:</u>
<u>MiniMax Lemma:</u> Let $f_i(x)$, $i = 1, ..., m$, be convex continuous functions on a convex compact set $X \subset \mathbb{R}^n$. Then there exists $\mu^* \geq 0$ with $\sum_i \mu_i^* = 1$ such that

$$\min_{x \in X} \max_{1 \leq i \leq m} f_i(x) = \min_{x \in X} \sum_i \mu_i^* f_i(x)$$

<u>Note:</u> When $\mu \geq 0, \sum_i \mu_i = 1$, one has

$$\begin{array}{rl} & \max\limits_{1 \leq i \leq m} f_i(x) \geq \sum\limits_i \mu_i f_i(x) \\ \Rightarrow & \min\limits_{x \in X} \max\limits_i f_i(x) \geq \min\limits_{x \in X} \sum\limits_i \mu_i f_i(x) \end{array}$$

**Proof of MinMax Lemma:** Consider the optimization program

$$\min_{t,x} \left\{ t : f_i(x) - t \leq 0,\, i \leq m, (t,x) \in X_+ \right\},$$
$$X_+ = \{(t,x) : x \in X\} \tag{P}$$

This program clearly is convex, solvable and satisfies the Slater condition, whence there exists $\lambda^* \geq 0$ and an optimal solution $(x_*, t_*)$ to $(P)$ such that $(x_*, \lambda^*)$ is the saddle point of the Lagrange function on $X^+ \times \{\lambda \geq 0\}$:

$$\begin{array}{rcll} \min\limits_{x \in X, t} \left\{ t + \sum\limits_i \lambda_i^*(f_i(x) - t) \right\} & = & t_* + \sum\limits_i \lambda_i^*(f_i(x_*) - t_*) & (a) \\ \max\limits_{\lambda \geq 0} \left\{ t + \sum\limits_i \lambda_i(f_i(x) - t) \right\} & = & t_* + \sum\limits_i \lambda_i^*(f_i(x_*) - t_*) & (b) \end{array}$$

$(b)$ implies that $t_* + \sum\limits_i \lambda_i^*(f_i(x_*) - t_*) = t_*$.

$(a)$ implies that $\sum\limits_i \lambda_i^* = 1$ and therefore implies that

$$\min_{x \in X} \sum_i \lambda_i^* f_i(x) = t_* = \min_{x \in X} \max_i f_i(x).$$

**Proof of Sion-Kakutani Theorem:** We should prove that problems

$$\begin{array}{rcll}
\mathrm{Opt}(P) & = & \inf\limits_{x\in X} \overbrace{\sup\limits_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} & (P)\\
\mathrm{Opt}(D) & = & \sup\limits_{\lambda\in\Lambda} \underbrace{\inf\limits_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} & (D)
\end{array}$$

are solvable with equal optimal values.

**1$^0$.** Since $X$ is compact and $F(x,\lambda)$ is continuous on $X\times\lambda$, the function $\underline{F}(\lambda)$ is continuous on $\Lambda$. Besides this, the sets

$$\Lambda^a = \{\lambda\in\Lambda : \underline{F}(\lambda)\geq a\}$$

are contained in the sets

$$\Lambda_a = \{\lambda\in\Lambda : F(\bar{x},\lambda)\geq a\}$$

and therefore are bounded. Finally, $\Lambda$ is closed, so that the *continuous* function $\underline{F}(\cdot)$ with *bounded* level sets $\Lambda^a$ attains it maximum on a *closed* set $\Lambda$. Thus, $(D)$ is solvable; let $\lambda^*$ be an optimal solution to $(D)$.

**$2^0$.** Consider the sets

$$X(\lambda) = \{x \in X : F(x, \lambda) \leq \mathrm{Opt}(D)\}.$$

These are closed convex subsets of a compact set $X$. Let us prove that every finite collection of these sets has a nonempty intersection. Indeed, assume that

$$X(\lambda^1) \cap ... \cap X(\lambda^N) = \emptyset.$$

so that

$$\max_{j=1,...,N} F(x, \lambda^j) > \mathrm{Opt}(D).$$

By MinMax Lemma, there exist weights $\mu_j \geq 0, \sum\limits_j \mu_j = 1$, such that

$$\min_{x \in X} \underbrace{\sum_j \mu_j F(x, \lambda^j)}_{\geq F(x, \underbrace{\sum_j \mu_j \lambda^j}_{\bar{\lambda}})} > \mathrm{Opt}(D)$$

which is impossible.

**3⁰.** Since every finite collection of closed convex subsets $X(\lambda)$ of a compact set has a nonempty intersection, all those sets have a nonempty intersection:

$$\exists x_* \in X : F(x_*, \lambda) \leq \mathrm{Opt}(D)\ \forall \lambda.$$

Due to $\mathrm{Opt}(P) \geq \mathrm{Opt}(D)$, this is possible iff $x_*$ is optimal for $(P)$ and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

# Optimality Conditions in Mathematical Programming

♣ Situation: We are given a Mathematical Programming problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \tag{P}$$

Question of interest: *Assume that we are given a feasible solution $x_*$ to $(P)$. What are the conditions (necessary, sufficient, necessary and sufficient) for $x_*$ to be optimal?*

Fact: Except for *convex programs*, there are no *verifiable local* sufficient conditions for *global* optimality. There exist, however,

◊verifiable local necessary conditions for *local* (and thus – for *global*) optimality

◊verifiable local sufficient conditions for *local* optimality

Fact: Existing conditions for local optimality assume that $x_* \in \text{int}X$, which, from the viewpoint of local optimality of $x_*$, is exactly the same as to say that $X = \mathbb{R}^n$.

♣ Situation: We are given a Mathematical Programming problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\}. \tag{P}$$

and a feasible solution $x_*$ to the problem, and are interested in necessary/sufficient conditions for *local* optimality of $x_*$:

There exists $r > 0$ such that for every feasible $x$ with $\|x - x_*\| \leq r$ one has

$$f(x) \geq f(x_*).$$

Default assumption: The objective and all the constraints are continuously differentiable in a neighbourhood of $x_*$.

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \le 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\}. \qquad (P)$$

♣ First Order Optimality Conditions are expressed via values and gradients of the objective and the constraints at $x_*$. Except for convex case, only necessary First Order conditions are known.

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\}. \qquad (P)$$

<u>The idea:</u>

♢Assume that $x_*$ is locally optimal for $(P)$. Let us approximate $(P)$ around $x_*$ by a Linear Programming program

$$\begin{array}{l} \min\limits_x f(x_*) + (x - x_*)^T f'(x_*) \\ \text{s.t.} \\ \overbrace{g_j(x_*)}^{0} + (x - x_*)^T g_j'(x_*) \leq 0,\ j \in J(x_*) \\ \underbrace{h_i(x_*)}_{0} + (x - x_*)^T h_i'(x_*) = 0,\ 1 \leq i \leq k \\ \left[ J(x_*) = \{j : g_j(x_*) = 0\} \right] \end{array} \qquad (LP)$$

<u>Note:</u> Since all $g_j(\cdot)$ are continuous at $x_*$, the *non-active at $x_*$* inequality constraints (those with $g_j(x_*) < 0$) do not affect $(LP)$.

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Rightarrow \min_x \left\{ (x - x_*)^T f'(x_*) : \begin{array}{l} (x - x_*)^T g_j'(x_*) \leq 0, \\ j \in J(x_*) \\ (x - x_*)^T h_i'(x_*) = 0, \\ i = 1, ..., k \end{array} \right\} \quad (LP)$$

$$J(x_*) = \{j : g_j(x_*) = 0\}$$

♢It is natural *to guess* that if $x_*$ is locally optimal for $(P)$, then $x_*$ is locally optimal for $(LP)$ as well.

LP is a *convex* program with *affine* constraints, whence the KKT conditions are necessary and sufficient for optimality:

$$x_* \text{ is optimal for } (LP)$$

$$\Updownarrow$$

$$\exists (\lambda_j^* \geq 0, j \in J(x_*), \mu_i) :$$

$$f'(x_*) + \sum_{j \in J(x_*)} \lambda_j^* g_j'(x_*) + \sum_{i=1}^k \mu_i h_i'(x_*) = 0$$

$$\Updownarrow$$

$$\exists (\lambda_j^* \geq 0, \mu_i^*) :$$

$$f'(x_*) + \sum_j \lambda_j^* g_j'(x_*) + \sum_i \mu_i^* h_i'(x_*) = 0$$

$$\lambda_j^* g_j(x_*) = 0,\ j = 1, ..., m$$

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \le 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \quad (P)$$

Proposition. Let $x_*$ be a locally optimal solution of $(P)$.

Assume that $x_*$ remains locally optimal when passing from $(P)$ to the linearized problem

$$\min_x \left\{ (x - x_*)^T f'(x_*) : \begin{array}{l} (x - x_*)^T g_j'(x_*) \le 0, \\ j \in J(x_*) \\ (x - x_*)^T h_i'(x_*) = 0, \\ i = 1, ..., k \end{array} \right\} \quad (LP)$$

Then at $x_*$ the KKT condition holds:

$$\begin{array}{l} \exists (\lambda_j^* \ge 0, \mu_i^*) : \\ \quad f'(x_*) + \sum_j \lambda_j^* g_j'(x_*) + \sum_i \mu_i^* h_i'(x_*) = 0 \\ \quad \lambda_j^* g_j(x_*) = 0, \; j = 1, ..., m \end{array}$$

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \le 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\min_x \left\{ (x - x_*)^T f'(x_*) : \begin{array}{l} (x - x_*)^T g_j'(x_*) \le 0, \\ j \in J(x_*) \\ (x - x_*)^T h_i'(x_*) = 0, \\ i = 1, ..., k \end{array} \right\} \quad (LP)$$

To make Proposition useful, we need a verifiable sufficient condition for "$x_*$ remains locally optimal when passing from $(P)$ to $(LP)$".

The most natural form of such a condition is *regularity*:

*Gradients, taken at $x_*$, of all constraints active at $x_*$ are linearly independent.*

Of course, all equality constraints by definition are active at every feasible solution.

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\min_x \left\{ (x - x_*)^T f'(x_*) : \begin{array}{l} (x - x_*)^T g_j'(x_*) \leq 0, \\ j \in J(x_*) \\ (x - x_*)^T h_i'(x_*) = 0, \\ i = 1, ..., k \end{array} \right\} \quad (LP)$$

Proposition: Let $x_*$ be a locally optimal *regular* solution of $(P)$. Then $x_*$ is optimal for $(LP)$ and, consequently, the KKT conditions take place at $x_*$.

Proof is based on an important fact of Analysis – a version of Implicit Function Theorem.

<u>Theorem:</u> Let $x_* \in \mathbb{R}^n$ and let $p_\ell(x)$, $\ell = 1, ..., L$, be real-valued functions such that

$\diamondsuit p_\ell$ are $\kappa \geq 1$ times continuously differentiable in a neighbourhood of $x_*$

$\diamondsuit p_\ell(x_*) = 0$

$\diamondsuit$vectors $\nabla p_\ell(x_*)$, $\ell = 1, ..., L$, are linearly independent.

Then there exists substitution of variables

$$y \mapsto x = \Phi(y)$$

defined in a neighbourhood $V$ of the origin and mapping $V$, in a one-to-one manner, onto a neighbourhood $B$ of $x_*$, such that

$\diamondsuit x_* = \Phi(0$

$\diamondsuit$both $\Phi : V \to B$ and its inverse mapping $\Phi^{-1} : B \to V$ are $\kappa$ times continuously differentiable

$\diamondsuit$in coordinates $y$, the functions $p_\ell$ become just the coordinates:

$$y \in V \Rightarrow p_\ell(\Phi(y)) \equiv y_\ell, \ell = 1, ..., L.$$

$$\min_x \left\{ (x-x_*)^T f'(x_*) : \begin{array}{l} (x-x_*)^T g_j'(x_*) \le 0, \\ j \in J(x_*) \\ (x-x_*)^T h_i'(x_*) = 0, \\ i = 1, ..., k \end{array} \right\} \quad (LP)$$

Let $x_*$ be a regular locally optimal solution to $(P)$; assume, on the contrary to what should be proven, that $x_*$ is not an optimal solution to $(LP)$, and let us lead this to contradiction.

**1**$^0$**.** Since $x = x_*$ is not an optimal solution to $(LP)$, there exists a feasible solution $x' = x_* + d$ to the problem with $(x' - x_*)^T f'(x_*) = d^T f'(x_*) < 0$, so that

$$d^T f'(x_*) < 0, \underbrace{d^T h_i'(x_*) = 0}_{\forall i}, \underbrace{d^T g_j'(x_*) \le 0}_{\forall j \in J(x_*)}$$

$$d^T f'(x_*) < 0, \underbrace{d^T h_i'(x_*) = 0}_{\forall i}, \underbrace{d^T g_j'(x_*) < 0}_{\forall j \in J(x_*)}$$

**$2^0$.** W.l.o.g., assume that $J(x_*) = \{1, ..., \ell\}$. By Theorem, there exist continuously differentiable local substitution of argument

$$x = \Phi(y) \qquad\qquad [\Phi(0) = x_*]$$

with a continuously differentiable in a neighbourhood of $x_*$ inverse $y = \Psi(x)$ such that in a neighbourhood of origin one has

$$h_i(\Phi(y)) \equiv y_i, \; g_j(\Phi(y)) = y_{k+j}, \, j = 1, ..., \ell.$$

Since $\Psi(\Phi(y)) \equiv y$, we have $\Psi'(x_*)\Phi'(0) = I$, whence

$$\exists e : \Phi'(0)e = d.$$

Situation: We have found a smooth local substitution of argument $x = \Phi(y)$ ($y = 0$ corresponds to $x = x_*$) and a direction $e$ such that in a neighbourhood of $y = 0$ one has

$$\begin{array}{llcl}
(a) & h_i(\Phi(y)) & \equiv & y_i,\ i \le k \\
(b) & g_j(\Phi(y)) & \equiv & y_{k+j},\ j \le \ell \\
 & & & [J(x_*) = \{1, ..., \ell\}] \\
(c) & [\Phi'(0)e]^T h_i'(x_*) & = & 0,\ i \le k \\
(d) & [\Phi'(0)e]^T g_j'(x_*) & < & 0,\ j \le \ell \\
(e) & [\Phi'(0)e]^T f'(x_*) & < & 0
\end{array}$$

Consider the differentiable curve

$$x(t) = \Phi(te).$$

We have

$$\begin{array}{l}
te_i \equiv h_i(\Phi(te)) \Rightarrow e_i = [\Phi'(0)e]^T h_i'(x_*) = 0 \\
te_{k+j} \equiv g_j(\Phi(te)) \Rightarrow e_{k+j} = [\Phi'(0)e]^T g_j'(x_*) < 0 \\
\Rightarrow \underbrace{h_i(x(t)) = te_i = 0}_{\forall i},\ \underbrace{g_j(x(t)) = te_{k+j} \le 0}_{\forall j \in J(x_*)}
\end{array}$$

Thus, $x(t)$ is feasible for all small $t \ge 0$. But:

$$\frac{d}{dt}\Big|_{t=0} f(x(t)) = [\Phi'(0)e]^T f'(x_*) < 0,$$

whence $f(x(t)) < f(x(0)) = f(x_*)$ for all small enough $t > 0$, which is a contradiction with local optimality of $x_*$.

## Second Order Optimality Conditions

♣ In the case of unconstrained minimization problem

$$\min_x f(x) \qquad (P)$$

with continuously differentiable objective, the KKT conditions reduce to Fermat Rule: *If* $x_*$ *is locally optimal for* $(P)$*, then* $\nabla f(x_*) = 0$.
Fermat Rule is the "first order" part of *Second Order Necessary Optimality Condition* in unconstrained minimization:
*If* $x_*$ *is locally optimal for* $(P)$ *and* $f$ *is twice differentiable in a neighbourhood of* $x_*$*, then*

$$\nabla f(x_*) = 0 \;\&\; \nabla^2 f(x_*) \succeq 0 \Leftrightarrow d^T \nabla^2 f(x_*) d \geq 0 \forall d$$

Indeed, let $x_*$ be locally optimal for $(P)$; then for appropriate $r_d > 0$

$$\begin{array}{l}
0 \leq t \leq r_d \\
\Rightarrow 0 \leq f(x_* + td) - f(x_*) \\
\quad = t\underbrace{d^T \nabla f(x_*)}_{=0} + \frac{1}{2}t^2 d^T \nabla^2 f(x_*) d + t^2 \underbrace{R_d(t)}_{\substack{\to 0, \\ t \to 0}} \\
\Rightarrow \frac{1}{2} d^T \nabla^2 f(x_*) d + R_d(t) \geq 0 \Rightarrow d^T \nabla^2 f(x_*) d \geq 0
\end{array}$$

$$\min_x f(x) \qquad (P)$$

The *necessary* Second Order Optimality condition in unconstrained minimization can be strengthened to

Second Order Sufficient Optimality Condition in unconstrained minimization: Let $f$ be twice differentiable in a neighbourhood of $x_*$. If

$$\nabla f(x_*) = 0, \nabla^2 f(x_*) \succ 0 \Leftrightarrow d^T \nabla^2 f(x_*) d > 0 \forall d \neq 0$$

then $x_*$ is locally optimal for $(P)$.

**Proof:** Since $d^T \nabla^2 f(x_*) d > 0$ for all $d > 0$, then there exists $\alpha > 0$ such that $d^T \nabla^2 f(x_*) d \geq \alpha d^T d$ for all $d$.

By differentiability, for every $\epsilon > 0$ there exists $r_\epsilon > 0$ such that

$$\begin{array}{l} \|d\|_2 \leq r_\epsilon \\ \Rightarrow f(x_* + d) - f(x_*) \geq \underbrace{d^T \nabla f(x_*)}_{=0} + \frac{1}{2} \underbrace{d^T \nabla^2 f(x_*) d}_{\geq \alpha d^T d} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \frac{\epsilon}{2} d^T d \\ \Rightarrow f(x_* + d) - f(x_*) \geq \frac{1}{2}(\alpha - \epsilon) d^T d \end{array}$$

Setting $\epsilon = \frac{\alpha}{2}$, we see that $x_*$ is a local minimizer of $f$.

We are given a Mathematical Programming problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Downarrow$$

$$L(x; \lambda, \mu) = f(x) + \sum_j \lambda_j g_j(x) + \sum_i \mu_i h_i(x)$$

♣ In Optimality Conditions for a constrained problem $(P)$, the role of $\nabla^2 f(x_*)$ is played by the Hessian of the *Lagrange function:*

Second Order Necessary Optimality Condition:

Let $x_*$ be a regular feasible solution of $(P)$ such that the functions $f, g_j, h_i$ are twice continuously differentiable in a neighbourhood of $x_*$. If $x_*$ is locally optimal, then

♢There exist uniquely defined Lagrange multipliers $\lambda_j^* \geq 0$, $\mu_i^*$ such that the KKT conditions hold:

$$\begin{array}{l} \nabla_x L(x_*; \lambda^*, \mu^*) = 0 \\ \lambda_j^* g_j(x_*) = 0,\ j = 1, ..., m \end{array}$$

♢For every $d$ orthogonal to the gradients, taken at $x_*$, of all active at $x_*$ equality and inequality constraints, one has

$$d^T \nabla_x^2 L(x_*; \lambda^*, \mu^*) d \geq 0.$$

**Proof. $1^0$.** Constraints which are non-active at $x_*$ clearly do not affect neither local optimality of $x_*$, nor the conclusion to be proven. Removing these constraints, we reduce the situation to one where *all* constraints in the problem

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \tag{$P$}$$

are active at $x_*$.

**$2^0$.** Applying Implicit Function Theorem, we can find a local change of variables

$$x = \Phi(y) \Leftrightarrow y = \Psi(x)$$
$$[\Phi(0) = x_*, \Psi(x_*) = 0]$$

with locally twice continuously differentiable $\Phi, \Psi$ such that

$$g_j(\Phi(y)) \equiv y_j,\ j \leq m, h_i(\Phi(y)) \equiv y_{m+i},\ i \leq k.$$

In variables $y$, problem $(P)$ becomes

$$\min_y \left\{ \underbrace{f(\Phi(y))}_{\phi(y)} : y_j \leq 0,\ j \leq m, y_{k+i} = 0,\ i \leq k \right\}. \tag{$P'$}$$

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Updownarrow \qquad [x = \Phi(y)]$$

$$\min_y \left\{ \underbrace{f(\Phi(y))}_{\phi(y)} : y_j \leq 0,\ j \leq m, y_{k+i} = 0,\ i \leq k \right\} \qquad (P')$$

$$\Downarrow$$

$$M(y; \lambda, \mu) = \phi(y) + \sum_j \lambda_j y_j + \sum_i \mu_i y_{m+i}$$

Our plan is as follows:

♢Since $\Phi$ is a smooth one-to-one mapping of a neighbourhood of $x_*$ onto a neighbourhood of $y_* = 0$, $x_*$ is locally optimal for $(P)$ iff $y_* = 0$ is locally optimal for $(P')$.

♢We intend to build necessary/sufficient conditions for $y_* = 0$ to be locally optimal for $(P')$; "translated" to $x$-variables, these conditions will imply necessary/sufficient conditions for local optimality of $x_*$ for $(P)$.

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Updownarrow \qquad [x = \Phi(y)]$$

$$\min_y \left\{ \underbrace{f(\Phi(y))}_{\phi(y)} : y_j \leq 0,\ j \leq m, y_{k+i} = 0,\ i \leq k \right\} \qquad (P')$$

$$\Downarrow$$

$$M(y; \lambda, \mu) = \phi(y) + \sum_j \lambda_j y_j + \sum_i \mu_i y_{m+i}$$

**3$^0$.** Since $x_* = \Phi(0)$ is locally optimal for $(P)$, $y_* = 0$ is locally optimal for $(P')$. In particular, if $e_i$ is $i$-th basic orth, then for appropriate $\epsilon > 0$:

$$\begin{array}{rcl} j \leq m & \Rightarrow & y(t) = -te_j \text{ is feasible for } (P') \text{ when} \\ \epsilon \geq t \geq 0 & \Rightarrow & -\frac{\partial \phi(0)}{\partial y_t} = \frac{d}{dt}\Big|_{t=0} \phi(y(t)) \geq 0 \\ & \Rightarrow & \lambda_j^* \equiv -\frac{\partial \phi(0)}{\partial y_i} \geq 0 \end{array}$$

and

$$\begin{array}{rcl} s > m + k & \Rightarrow & y(t) = te_s \text{ is feasible for } (P') \text{ when} \\ \epsilon \geq t \geq \epsilon & \Rightarrow & \frac{\partial \phi(0)}{\partial y_s} = \frac{d}{dt}\Big|_{t=0} \phi(y(t)) = 0 \end{array}$$

Setting $\mu_i^* = -\frac{\partial \phi(0)}{\partial y_{m+i}}$, $i = 1, ..., k$, we get

$$\lambda^* \geq 0 \ \&\ \nabla_y M(0; \lambda^*, \mu^*) = 0. \qquad \text{(KKT)}$$

Situation: $y_* = 0$ is locally optimal for

$$\min_y \left\{ \phi(y) \equiv f(\Phi(y)) : \begin{array}{l} y_j \leq 0,\ j \leq m \\ y_m + i = 0,\ i \leq k \end{array} \right\} \quad (P')$$

$$\Downarrow$$

$$M(y;\lambda,\mu) = \phi(y) + \sum_{j=1}^{m} \lambda_j y_j + \sum_{i=1}^{k} \mu_i y_{m+i}$$

and $\exists \lambda^* \geq 0, \mu^*$:

$$0 = \frac{\partial M(0;\lambda^*,\mu^*)}{\partial y_\ell} \equiv \begin{cases} \frac{\partial \phi(0)}{\partial y_\ell} + \lambda_j^*, & \ell \leq m \\ \frac{\partial \phi(0)}{\partial y_\ell} + \mu_{\ell-m}^*, & m < \ell \leq m+k \\ \frac{\partial \phi(0)}{\partial y_\ell}, & \ell > m+k \end{cases}$$

(KKT)

Note that the condition $\nabla_y M(0;\lambda^*,\mu^*) = 0$ defines $\lambda^*$, $\mu^*$ are in a unique fashion.

**$4^0$.** We have seen that for $(P')$, the first order part of the Necessary Second Order Optimality condition holds true. Let us prove the second order part of the condition, which reads

$$\begin{array}{c} \forall (d : d^T \nabla_y y_\ell = 0,\ \ell \leq m+k) : \\ d^T \nabla_y^2 M(0;\lambda^*,\mu^*) d \geq 0. \end{array}$$

Situation: $y_* = 0$ is locally optimal solution to the problem

$$\min_y \left\{ \phi(y) \equiv f(\Phi(y)) : \begin{array}{l} y_j \leq 0,\, j \leq m \\ y_m + i = 0,\, i \leq k \end{array} \right\} \qquad (P')$$

Claim:

$$\forall (d : d^T \nabla_y y_\ell = 0,\, \ell \leq m+k) : \\ d^T \nabla_y^2 M(0; \lambda^*, \mu^*) d \geq 0.$$

This is evident: since $M(y; \lambda^*, \mu^*) = \phi(y) + \sum_{j=1}^{m} \lambda_j^* y_j + \sum_{i=1}^{k} \mu_i^* y_{m+i}$, we have

$$\nabla_y^2 M(0; \lambda^*, \mu^*) = \nabla^2 \phi(0).$$

Claim therefore states that $d^T \nabla^2 \phi(0) d \geq 0$ for every vector $d$ from the linear subspace $L = \{d : d_1 = \ldots = d_{m+k} = 0\}$. But this subspace is feasible for $(P')$, so that $\phi$, restricted onto $L$, should attain unconstrained local minimum at the origin. By Necessary Second Order Optimality condition for unconstrained minimization,

$$d^T \nabla^2 \phi(0) d \geq 0 \;\forall d \in L.$$

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Updownarrow \qquad [x = \Phi(y)]$$

$$\min_y \left\{ \underbrace{f(\Phi(y))}_{\phi(y)} : y_j \leq 0,\ j \leq m, y_{k+i} = 0,\ i \leq k \right\} \qquad (P')$$

$$\Downarrow$$

$$M(y; \lambda, \mu) = \phi(y) + \sum_j \lambda_j y_j + \sum_i \mu_i y_{m+i}$$

**$5^0$.** We have seen that if $x_*$ is locally optimal for $(P)$, then there exist uniquely defined $\lambda^* \geq 0$, $\mu^*$ such that

$$\nabla_y M(0; \lambda^*, \mu^*) = 0,$$

and one has

$$d^T \nabla_y y_\ell = 0,\ \ell \leq m+k \Rightarrow d^T \nabla_y^2 M(0; \lambda^*, \mu^*) d \geq 0.$$

Let us prove that then

$$\nabla_x L(x_*; \lambda^*, \mu^*) = 0 \qquad (*)$$

and

$$\left. \begin{array}{l} e^T g_j'(x_*) = 0,\ j \leq m \\ e^T h_i'(x_*) = 0,\ i \leq k \end{array} \right\} \Rightarrow e^T \nabla_x^2 L(x_*); \lambda^*, \mu^*) e \geq 0. \qquad (**)$$

Given:

$$\nabla_y M(0;\lambda^*,\mu^*) = 0$$
$$d^T\nabla_y y_\ell = 0,\ \ell \le m+k \Rightarrow d^T\nabla_y^2 M(0;\lambda^*,\mu^*)d \ge 0.$$

Should prove:

$$\nabla_x L(x_*;\lambda^*,\mu^*) = 0 \qquad (*)$$
$$\left.\begin{array}{l} e^T g_j'(x_*) = 0,\ j \le m \\ e^T h_i'(x_*) = 0,\ i \le k \end{array}\right\} \Rightarrow e^T\nabla_x^2 L(x_*);\lambda^*,\mu^*)e \ge 0 \qquad (**)$$

♢Setting $\mathcal{L}(x) = L(x;\lambda^*,\mu^*)$, $\mathcal{M}(y) = M(y;\lambda^*,\mu^*)$, we have

$$\begin{array}{rl} & \mathcal{L}(x) = \mathcal{M}(\Psi(x)) \\ \Rightarrow & \nabla_x\mathcal{L}(x_*) = [\Psi'(x_*)]^T\nabla_y\mathcal{M}(y_*) = 0, \end{array}$$

as required in $(*)$.

♢Let $e$ satisfy the premise in $(**)$, and let $d = [\Phi'(0)]^{-1}e$. Then

$$\overbrace{\left.\frac{d}{dt}\right|_{t=0} g_j(\Phi(td))}^{\left.\frac{d}{dt}\right|_{t=0} td_j} = [g_j'(x_*)]^T\overbrace{[\Phi'(0)]d}^{e}$$
$$\Rightarrow d_j = e^T g_j'(x_*) = 0,\ j \le m$$
$$\underbrace{\left.\frac{d}{dt}\right|_{t=0} h_i(\Phi(td))}_{\left.\frac{d}{dt}\right|_{t=0} td_{m+i}} = [h_i'(x_*)]^T\underbrace{[\Phi'(0)]d}_{e}$$
$$\Rightarrow d_{m+i} = e^T h_i'(x_*) = 0,\ i \le k$$

We have

$$
\begin{array}{rl}
& e^T\nabla^2\mathcal{L}(x_*)e = \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{L}(x_*+te) = \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{M}(\Psi(x_*+te)) \\
= & \frac{d}{dt}\Big|_{t=0}\left[e^T[\Psi'(x_*+te)]^T\nabla\mathcal{M}(\Psi(x_*+te))\right] \\
= & e^T[\Psi'(x_*)]^T\nabla^2\mathcal{M}(0)[\Psi'(x_*)e] \\
& +e^T[\frac{d}{dt}_{t=0}\Psi'(x_*+te)]^T\underbrace{\nabla\mathcal{M}(0)}_{=0} \\
= & d^T\nabla^2\mathcal{M}d \geq 0,
\end{array}
$$

Thus, whenever $e$ is orthogonal to the gradients of all constraints active at $x_*$, we have $e^T\nabla^2\mathcal{L}e \geq 0$.

Second Order Sufficient Condition for Local Optimality

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \qquad (P)$$

$$\Downarrow$$

$$L(x;\lambda,\mu) = f(x) + \sum_j \lambda_j g_j(x) + \sum_i \mu_i h_i(x)$$

Second Order Sufficient Optimality Condition:

Let $x_*$ be a regular feasible solution of $(P)$ such that the functions $f, g_j, h_i$ are twice continuously differentiable in a neighbourhood of $x_*$. If there exist Lagrange multipliers $\lambda_j^* \geq 0$, $\mu_i^*$ such that

$\diamondsuit$the KKT conditions hold:

$$\begin{array}{l} \nabla_x L(x_*; \lambda^*, \mu^*) = 0 \\ \lambda_j^* g_j(x_*) = 0,\ j = 1, ..., m \end{array}$$

$\diamondsuit$For every $d \neq 0$ orthogonal to the gradients, taken at $x_*$, of all active at $x_*$ equality constraints and those active at $x_*$ inequality constraints for which $\lambda_j^* > 0$, one has

$$d^T \nabla_x^2 L(x_*; \lambda^*, \mu^*) d > 0$$

then $x_*$ is locally optimal for $(P)$.

Note: Difference between Sufficient and Necessary optimality conditions is in their "second order" parts and is twofold:

♢[minor difference] Necessary condition states positive semidefiniteness of $\nabla_x^2 L(x_*;\lambda^*,\mu^*)$ along linear subspace:

$$\forall d \in T = \{d : \overbrace{d^T h_i'(x_*) = 0}^{\forall i \le k}, \overbrace{d^T g_j'(x_*) = 0}^{\forall j \in J(x_*)}\} : \\ d^T \nabla_x^2 L(x_*;\lambda^*,\mu^*) d \ge 0$$

while Sufficient condition requires positive definiteness of $\nabla_x^2 L(x_*;\lambda^*,\mu^*)$ along linear subspace:

$$\forall 0 \ne d \in T^+ = \{d : \overbrace{d^T h_i'(x_*) = 0}^{\forall i \le k}, \overbrace{d^T g_j'(x_*) = 0}^{\forall j : \lambda_j^* > 0}\} : \\ d^T \nabla_x^2 L(x_*;\lambda^*,\mu^*) d > 0$$

♢[major difference] The linear subspaces in question are different, and $T \subset T^+$; the subspaces are equal to each other iff *all* active at $x_*$ inequality constraints have positive Lagrange multipliers $\lambda_j^*$.

Note: This "gap" is essential, as is shown by example

$$\min_{x_1,x_2}\left\{f(x)=x_2^2-x_1^2 : g_1(x)=x_1\leq 0\right\}$$

$$[x_*=(0,0)^T]$$

Here the Necessary Second Order Optimality condition is satisfied "strictly": $L(x;\lambda)=x_2^2-x_1^2+\lambda x_1$, whence

$$\begin{array}{c}\lambda^*=0\Rightarrow\nabla_x L(x_*;\lambda^*)=0,\\ T=\{d:d^Tg_1'(0)=0\}=\{d:d_1=0\},\\ 0\neq d\in T\Rightarrow d^T\nabla_x^2L(x_*;\lambda^*)d=d_2^2>0\}\end{array}$$

while $x_*$ is not a local solution.

**Proof of Sufficient Second Order Optimality Condition. $1^0$.** As in the case of Second Order Necessary Optimality Condition, we can reduce the situation to one where

♢All inequality constraints are active at $x_*$

♢The problem is of the special form

$$\min_y \left\{ \phi(y) : \begin{array}{l} y_j \le 0,\ j \le m \\ y_{m+i} = 0,\ i \le k \end{array} \right\} \qquad (P')$$

**$2^0$.** In the case of $(P')$, Sufficient condition reads: $\exists \lambda^* \ge 0, \mu^*$:

$$\begin{array}{c} \nabla_y\big|_{y=0} \left\{ \phi(y) + \sum_{j=1}^m \lambda_j^* y_j + \sum_{i=1}^k \mu_i^* y_{m+i} \right\} \\ d_j = 0, j \in J, d \ne 0 \Rightarrow d^T \nabla^2 \phi(0) d > 0 \\ \left[ J = \{j \le m : \lambda_j^* > 0\} \cup \{m+1, ..., m+k\} \right] \end{array} \qquad (*)$$

Assuming w.l.o.g. $\{j : \lambda_j^* > 0\} = \{1, ..., q\}$, $(*)$ reads:

$$\begin{array}{rcl} \frac{\partial \phi(0)}{\partial y_\ell} & < & 0,\ \ell = 1, ..., q \\ \frac{\partial \phi(0)}{\partial y_\ell} & = & 0,\ \ell = q+1, ..., m \\ \frac{\partial \phi(0)}{\partial y_\ell} & = & 0,\ \ell = m+k+1, ..., n \\ 0 \ne d & \in & T^+ = \{d : d_\ell = 0,\ \ell \in \{1, ..., q, m+1, ..., m+k\}\} : \\ & \Rightarrow & d^T \nabla^2 \phi(0) d > 0 \end{array}$$

Our goal is to derive from this assumption local optimality of $y_* = 0$ for $(P')$.

**$2^0$.** The feasible set of $(P')$ is the closed cone

$K = \{d : d_\ell \le 0,\ \ell = 1, ..., m, d_\ell = 0,\ \ell = m+1, ..., m+k\}$

Lemma: *For* $0 \ne d \in K$ *one has* $d^T \nabla \phi(0) \ge 0$ *and*

$$d^T \nabla \phi(0) = 0 \Rightarrow d^T \nabla^2 \phi(0) d > 0.$$

Situation:

$$\begin{array}{rcl}
\frac{\partial\phi(0)}{\partial y_\ell} & < & 0,\ \ell = 1,...,q \\
\frac{\partial\phi(0)}{\partial y_\ell} & = & 0,\ \ell = q+1,...,m \\
\frac{\partial\phi(0)}{\partial y_\ell} & = & 0,\ \ell = m+k+1,...,n \\
0 \neq d & \in & T^+ = \{d : d_\ell = 0,\ \ell \in \{1,...,q,m+1,...,m+k\}\} : \\
& \Rightarrow & d^T\nabla^2\phi(0)d > 0
\end{array} \qquad (*)$$

$$K = \{d : d_\ell \le 0,\ \ell = 1,...,m, d_\ell = 0,\ \ell = m+1,...,m+k\} \qquad (**)$$

Claim: *For $0 \neq d \in K$ one has $d^T\nabla\phi(0) \geq 0$ and*

$$d^T\nabla\phi(0) = 0 \Rightarrow d^T\nabla^2\phi(0)d > 0.$$

**Proof:** For $d \in K$, we have

$$d^T\nabla\phi(0) = \sum_{\ell=1}^{n} \frac{\partial\phi(0)}{\partial y_\ell} d_\ell$$

By $(*) - (**)$, the first $q$ terms in this sum are nonnegative, and the remaining are 0. Thus, the sum always is $\geq 0$. For $d \neq 0$, the only possibility for the sum to vanish is to have $d \in T^+$, and in this case $d^T\phi''(0)d > 0$.

Situation: $(P')$ is the problem

$$\min_{y \in K} \phi(y), \tag{!}$$

$K$ is a closed cone, $\phi$ is twice continuously differentiable in a neighbourhood of the origin and is such that

$$\begin{array}{c} d \in K \Rightarrow d^T \nabla \phi(0) \geq 0 \\ d \in K \backslash \{0\}, d^T \nabla \phi(0) = 0 \Rightarrow d^T \nabla^2 \phi(0) d > 0 \end{array}$$

Claim: *In the situation in question, 0 is a locally optimal solution to (!).*

**Proof:** Let $M = \{d \in K : \|d\|_2 = 1\}$, and let $M_0 = \{d \in M : d^T \nabla \phi(0) = 0\}$. Since $K$ is closed, both $M$ and $M_0$ are compact sets. We know that $d^T \nabla^2 \phi(0) d > 0$ for $d \in M_0$. Since $M_0$ is a compact set, there exists a neighbourhood $V$ of $M_0$ and $\alpha > 0$ such that

$$d \in V \Rightarrow d^T \nabla^2 \phi(0) d \geq \alpha.$$

The set $V_1 = M \backslash V$ is compact and $d^T \nabla \phi(0) > 0$ when $d \in V_1$; thus, there exists $\beta > 0$ such that

$$d \in V_1 \Rightarrow d^T \nabla \phi(0) \geq \beta.$$

Situation: $K$ is a cone, and the set $M = \{d \in K : \|d\|_2 = 1\}$ is partitioned into two subsets $V_0 = V \cap M$ and $V_1$ in such a way that

$$\begin{array}{rcl} d \in V_0 & \Rightarrow & d^T \nabla\phi(0) \geq 0, d^T \nabla^2 \phi(0) d \geq \alpha > 0 \\ d \in V_1 & \rightarrow & d^T \nabla\phi(0) \geq \beta > 0 \end{array}$$

Goal: To prove that 0 is local minimizer of $\phi$ on $K$, or, which is the same, that

$$\begin{array}{l} \exists r > 0 : \\ \qquad \phi(0) \;\leq\; \phi(td) \; \forall (d \in M, 0 \leq t \leq r). \end{array}$$

**Proof:** Let $d \in M$, $t \geq 0$. When $d \in V_0$, we have

$$\begin{array}{rcl} \phi(td) - \phi(0) & \geq & td^T \nabla\phi(0) + \frac{1}{2} t^2 d^T \nabla^2 \phi(0) d - t^2 \underbrace{R(t)}_{\substack{\to 0, \\ t \to +0}} \\ & \geq & \frac{1}{2} t^2 (\alpha - 2R(t)) \\ \Rightarrow \exists r_0 > 0 : & & \phi(td) - \phi(0) \geq \frac{1}{4} t^2 \alpha \geq 0 \, \forall t \leq r_0 \end{array}$$

When $d \in V_1$, we have

$$\begin{array}{rcl} \phi(td) - \phi(0) & \geq & td^T \nabla\phi(0) + \frac{1}{2} t^2 d^T \nabla^2 \phi(0) d - t^2 \underbrace{R(t)}_{\substack{\to 0, \\ t \to +0}} \\ & \geq & \beta t - C t^2 - t^2 R(t) \\ \Rightarrow \exists r_1 > 0 : & & \phi(td) - \phi(0) \geq \frac{\beta}{2} t \geq 0 \, \forall t \leq r_1 \end{array}$$

Thus, $\phi(td) - \phi(0) \geq 0$ for all $t \leq \min[r_0, r_1]$, $d \in M$.

# Sensitivity Analysis

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \le 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \quad (P)$$

$$\Downarrow$$

$$L(x; \lambda, \mu) = f(x) + \sum_j \lambda_j g_j(x) + \sum_i \mu_i h_i(x)$$

Definition: Let $x_*$ be a feasible solution to $(P)$ such that the functions $f, g_j, h_i$ are $\ell \ge 2$ times continuously differentiable in a neighbourhood of $x_*$.

$x_*$ is called a *nondegenerate locally optimal* solution to $(P)$, if

♢ $x_*$ is a regular solution (i.e., gradients of active at $x_*$ constraints are linearly independent)

♢ at $x_*$, Sufficient Second Order Optimality condition holds $\exists (\lambda^* \ge 0, \mu^*)$:

$$\begin{array}{l} \nabla_x L(x_*; \lambda^*, \mu^*) = 0 \\ \lambda_j^* g_j(x_*) = 0,\ j = 1, ..., m \\ \left. \begin{array}{l} d^T \nabla g_j(x_*) = 0\, \forall (j : \lambda_j^* > 0) \\ d^T \nabla h_i(x_*) = 0\, \forall i \\ d \ne 0 \end{array} \right\} \Rightarrow d^T \nabla_x^2 L(x_*; \lambda^*, \mu^*) > 0 \end{array}$$

♢ for all active at $x_*$ inequality constraints, Lagrange multipliers are positive:

$$g_j(x_*) = 0 \Rightarrow \lambda_j^* > 0.$$

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \end{array} \right\} \quad (P)$$

Theorem: Let $x_*$ be a nondegenerate locally optimal solution to $(P)$. Let us embed $(P)$ into the parametric family of problems

$$\min_x \left\{ f(x) : \begin{array}{c} g_1(x) \leq a_1, ..., g_m(x) \leq a_m \\ h_1(x) = b_1, ..., h_k(x) = b_k \end{array} \right\} \qquad (P[a,b])$$

so that $(P)$ is $(P[0,0])$.
There exists a neighbourhood $V_x$ of $x_*$ and a neighbourhood $V_{a,b}$ of the point $a = 0, b = 0$ in the space of parameters $a, b$ such that
$\diamondsuit \forall (a,b) \in V_{a,b}$, in $V_v$ there exists a unique KKT point $x_*(a,b)$ of $(P[a,b])$, and this point is a nondegenerate locally optimal solution to $(P[a,b])$; moreover, $x_*(a,b)$ is optimal solution for the optimization problem

$$\mathrm{Opt}_{\mathrm{loc}}(a,b) = \min_x \left\{ f(x) : \begin{array}{c} g_1(x) \leq a_1, ..., g_m(x) \leq a_m \\ h_1(x) = b_1, ..., h_k(x) = b_k \\ x \in V_x \end{array} \right\}$$
$$(P_{\mathrm{loc}}[a,b])$$

$\diamondsuit$both $x_*(a,b)$ and the corresponding Lagrange multipliers $\lambda^*(a,b)$, $\mu^*(a,b)$ are $\ell-1$ times continuously differentiable functions of $(a,b)\in V_{a,b}$, and

$$\begin{array}{l}\frac{\partial \mathrm{Opt}_{\mathrm{loc}}(a,b)}{\partial a_j}=\frac{\partial f(x_*(a,b))}{\partial a_j}=-\lambda_j^*(a,b)\\ \frac{\partial \mathrm{Opt}_{\mathrm{loc}}(a,b)}{\partial b_i}=\frac{\partial f(x_*(a,b))}{\partial b_i}=-\mu_i^*(a,b)\end{array}$$

## Simple example: Existence of Eigenvalue

♣ Consider optimization problem

$$\mathrm{Opt} = \min_{x\in\mathbb{R}^n} \left\{ f(x) = x^T A x : h(x) = 1 - x^T x = 0 \right\} \qquad (P)$$

where $A = A^T$ is an $n \times n$ matrix. The problem clearly is solvable. Let $x_*$ be its optimal solution. What can we say about $x_*$?

Claim: $x_*$ is a regular solution to $(P)$.

Indeed, we should prove that the gradients of active at $x_*$ constraints are linearly independent. There is only one constraint, and its gradient at the feasible set is nonzero.

Since $x_*$ is a regular globally (and therefore locally) optimal solution, at $x_*$ the Necessary Second Order Optimality condition should hold: $\exists \mu^*$:

$$\nabla_x \overbrace{\left[x^T A x + \mu^*(1 - x^T x)\right]}^{L(x;\mu^*)} = 0 \Leftrightarrow 2(A - \mu^* I)x_* = 0$$

$$\underbrace{d^T \nabla_x h(x_*) = 0}_{\Leftrightarrow d^T x_* = 0} \Rightarrow \underbrace{d^T \nabla_x^2 L(x_*; \mu^*) d \geq 0}_{\Leftrightarrow d^T (A - \mu^* I) d \geq 0}$$

$$\text{Opt} = \min_{x\in\mathbb{R}^n} \left\{ f(x) = x^T A x : g(x) = 1 - x^T x = 0 \right\} \qquad (P)$$

Situation: If $x_*$ is optimal, then $\exists \mu^*$:

$$\begin{array}{rr} Ax_* = \mu^* x_* & (A) \\ d^T x_* = 0 \Rightarrow d^T(A - \mu^* I)d \geq 0 & (B) \end{array}$$

♣ $(A)$ says that $x_* \neq 0$ is an eigenvector of $A$ with eigenvalue $\mu^*$; in particular, *we see that a symmetric matrix always has a real eigenvector*

♣ $(B)$ along with $(A)$ says that $y^T(A-\mu^* I)y \geq 0$ *for all* $y$. Indeed, every $y \in \mathbb{R}^n$ can be represented as $y = tx_* + d$ with $d^T x_* = 0$. We now have

$$\begin{array}{rl} & y^T[A - \mu^* I]y = (tx_* + d)^T[A - \mu^* I](tx_* + d) \\ = & t^2 x_*^T \underbrace{[A - \mu^* I]x_*}_{=0} + 2td^T \underbrace{d^T[A - \mu^* I]x_*}_{=0} \\ & + \underbrace{d^T[A - \mu^* I]d}_{\geq 0} \geq 0 \end{array}$$

$$\text{Opt} = \min_{x\in\mathbb{R}^n} \left\{ f(x) = x^T A x : g(x) = 1 - x^T x = 0 \right\} \qquad (P)$$

Note: In the case in question, Necessary Second Order Optimality conditions can be rewritten equivalently as $\exists \mu^*$:

$$\begin{array}{c} [A - \mu^* I] x_* = 0 \\ y^T [A - \mu^* I] y \geq 0 \,\forall y \end{array} \qquad (*)$$

and are not only necessary, but also sufficient for feasible solution $x_*$ to be globally optimal.

To prove sufficiency, let $x_*$ be feasible, and $\mu^*$ be such that $(*)$ holds true. For every feasible solution $x$, one has

$$0 \leq x^T [A - \mu^* I] x = x^T A x - \mu^* x^T x = x^T A x - \mu^*,$$

whence $x^T A x \geq \mu^*$. For $x = x_*$, we have

$$0 = x_*^T [A - \mu^* I] x_* = x_*^T A x_* - \mu^* x_*^T x_* = x_*^T A x_* - \mu^*,$$

whence $x_*^T A x_* = \mu^*$. Thus, $x_*$ is globally optimal for $(P)$, and $\mu^*$ is the optimal value in $(P)$.

Extension: $\mathcal{S}$-Lemma. Let $A, B$ be symmetric matrices, and let $B$ be such that

$$\exists \bar{x} : \bar{x}^T B \bar{x} > 0. \qquad (*)$$

Then the inequality

$$x^T A x \geq 0 \qquad (A)$$

is a consequence of the inequality

$$x^T B x \geq 0 \qquad (B)$$

iff $(A)$ is a "linear consequence" of $(B)$: there exists $\lambda \geq 0$ such that

$$x^T [A - \lambda B] x \geq 0 \forall x \qquad (C)$$

that is, $(A)$ is a weighted sum of $(B)$ (weight $\lambda \geq 0$) and identically true inequality $(C)$.

Sketch of the proof: The only nontrivial statement is that "If $(A)$ is a consequence of $(B)$, then t e exists $\lambda \geq 0$ such that ...". To prove this statement, assume that $(A)$ is a consequence of $(B)$.

Situation:

$$\exists \bar{x} : \bar{x}^T B \bar{x} > 0;\ \underbrace{x^T B x \geq 0}_{(B)} \Rightarrow \underbrace{x^T A x \geq 0}_{(A)}$$

Consider optimization problem

$$\mathrm{Opt} = \min_x \left\{ x^T A x : h(x) \equiv 1 - x^T B x = 0 \right\}.$$

Problem is feasible by $(*)$, and Opt $\geq 0$. Assume that an optimal solution $x_*$ exists. Then, same as above, $x_*$ is regular, and at $x_*$ the Second Order Necessary condition holds true: $\exists \mu^*$:

$$\nabla_x\big|_{x=x_*} \left[ x^T A x + \mu^* [1 - x^T B x] \right] = 0 \Leftrightarrow [A - \mu^* B] x_* = 0$$
$$\underbrace{d^T \nabla_x\big|_{x=x_*} h(x) = 0}_{\Leftrightarrow d^T B x_* = 0} \Rightarrow d^T [A - \mu^* B] d \geq 0$$

We have $0 = x_*^T [A - \mu^* B] x_*$, that is, $\mu_* = \mathrm{Opt} \geq 0$. Representing $y \in \mathbb{R}^n$ as $t x_* + d$ with $d^T B x_* = 0$ (that is, $t = x_*^T B y$), we get

$$y^T [A - \mu^* B] y = t^2 x_*^T \underbrace{[A - \mu^* B] x_*}_{=0}$$
$$+ 2 t d^T \underbrace{[A - \mu^* B] x_*}_{=0} + \underbrace{d^T [A - \mu^* B] d}_{\geq 0} \geq 0,$$

Thus, $\mu^* \geq 0$ and $y^T [A - \mu^* B] y \geq 0$ for all $y$, Q.E.D.

Introduction to Optimization Algorithms

♣ Goal: Approximate numerically solutions to Mathematical Programming problems

$$\min_x \left\{ f(x) : \begin{array}{l} g_j(x) \leq 0,\, j = 1, ..., m \\ h_i(x) = 0,\, i = 1, ..., k \end{array} \right\} \qquad (P)$$

♣ Traditional MP algorithms to be considered in the Course do *not* assume the analytic structure of $(P)$ to be known in advance (and do not know how to use the structure when it is known). These algorithms are *black-box-oriented*: when solving $(P)$, method generates a sequence of *iterates* $x_1$, $x_2$,... in such a way that $x_{t+1}$ *depends solely on local information of* $(P)$ *gathered along the preceding iterates* $x_1, ..., x_t$.

Information on $(P)$ obtained at $x_t$ usually is comprised of the values and the first and the second derivatives of the objective and the constraints at $x_t$.

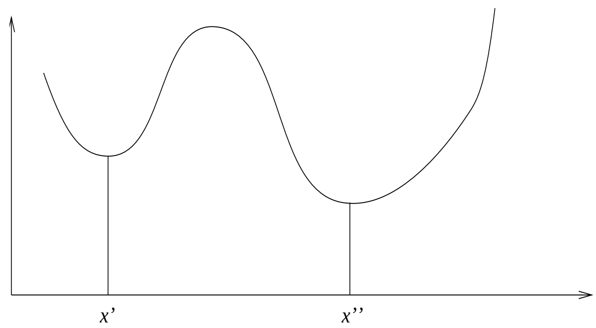
♠ Note: In optimization, there exist algorithms which do exploit problem's structure. Traditional methods of this type – Simplex method and its variations – do not go beyond Linear Programming and Linearly Constrained Quadratic Programming.

Recently, new efficient ways to exploit problem's structure were discovered (Interior Point methods). The resulting algorithms, however, do not go beyond Convex Programming.

♣ *Except for very specific and relatively simple problem classes*, like Linear Programming or Linearly Constrained Quadratic Programming, optimization algorithms cannot guarantee finding exact solution – local or global – in finite time. The best we can expect from these algorithms is *convergence* of approximate solutions generated by algorithms to the exact solutions.

♠ Even in the case when "finite" solution methods do exist (Simplex method in Linear Programming), no reasonable complexity bounds for these methods are known, therefore in reality the ability of a method to generate the exact solution in finitely many steps is neither necessary, nor sufficient to justify the method.

♣ *Aside of Convex Programming*, traditional optimization methods are unable to *guarantee* convergence to a *globally optimal* solution. Indeed, in the non-convex case there is no way to conclude from *local* information whether a given point is/is not globally optimal:




"looking" at problem around $x'$, we get absolutely no hint that the trues global optimal solution is $x''$.

♠ In order to *guarantee* approximating *global* solution, it seems unavoidable to "scan" a dense set of the values of $x$ in order to be sure that the globally optimal solution is not missed. Theoretically, such a possibility exists; however, the complexity of "exhaustive search" methods blows up exponentially with the dimension of the decision vector, which makes these methods completely impractical.

♣ Traditional optimization methods do *not* incorporate exhaustive search and, as a result, can*not* guarantee convergence to a global solution.

♠ A typical theoretical result on a traditional the optimization method as applied to a general (not necessary convex) problem sounds like:

> *Assume that problem $(P)$ possesses the following properties:*
> *...*
> *...*
> *Then the sequence of approximate solutions generated by method X is bounded, and all its limiting points are KKT points of the problem.*

or

> *Assume that $x_*$ is a nondegenerate local solution to $(P)$. Then method X, started close enough to $x_*$, converges to $x_*$.*

## Classification of MP Algorithms

♣ There are two major traditional classifications of MP algorithms:

♢*Classification by application fields,* primarily into

- algorithms for unconstrained optimization
- algorithms for constrained optimization

♢*Classification by information* used by the algorithms, primarily into

- zero order methods which use only the values of the objective and the constraints
- first order methods (use both values and first order derivatives)
- second order methods (use values, first- and second order derivatives).

## Rate of Convergence of MP Algorithms

♣ There is a necessity to quantify the convergence properties of MP algorithms. Traditionally, this is done via *asymptotical rate of convergence* defined as follows:

Step 1. We introduce an appropriate *error measure* – a nonnegative function $\mathrm{Error}_P(x)$ of approximate solution and of the problem we are solving which is zero exactly at the set $X_*$ of solutions to $(P)$ we intend to approximate.

Examples: (i) Distance to the set $X_*$:

$$\mathrm{Error}_P(x) = \inf_{x_* \in X_*} \|x - x_*\|_2$$

(ii) Residual in terms of the objective and the constraints

$$\mathrm{Error}_P(x) = \max \Big[ f(x) - \mathrm{Opt}(P), \\ [g_1(x)]_+, ..., [g_m(x)]_+, \\ |h_1(x)|, ..., |h_k(x)| \Big]$$

Step 2. Assume that we have established *convergence* of our method, that is, we know that if $x_t^*$ are approximate solutions generated by method as applied to a problem $(P)$ from a given family, then

$$\text{Error}_P(t) \equiv \text{Error}_P(x_t^*) \to 0,\ t \to \infty$$

We then roughly quantify the *rate* at which the sequence $\text{Error}_P(t)$ of nonnegative reals converges to 0. Specifically, we say that

♢the method converges *sublinearly*, if the error goes to zero less rapidly than a geometric progression, e.g., as $1/t$ or $1/t^2$;

♢the method converges *linearly*, if there exist $C < \infty$ and $q \in (0,1)$ such that

$$\text{Error}_{(P)}(t) \leq Cq^t$$

$q$ is called the *convergence ratio.* E.g.,

$$\text{Error}_P(t) \asymp e^{-at}$$

exhibits linear convergence with ratio $e^{-a}$.

Sufficient condition for linear convergence with ratio $q \in (0,1)$ is that

$$\overline{\lim_{t\to\infty}} \frac{\text{Error}_P(t+1)}{\text{Error}_P(t)} < q$$

$\diamondsuit$the method converges *superlinearly*, if the sequence of errors converges to 0 faster than every geometric progression:

$$\forall q \in (0,1) \exists C : \text{Error}_P(t) \leq C q^t$$

For example,

$$\text{Error}_P(t) \asymp e^{-at^2}$$

corresponds to superlinear convergence.
Sufficient condition for superlinear convergence is

$$\lim_{t\to\infty} \frac{\text{Error}_P(t+1)}{\text{Error}_P(t)} = 0$$

$\diamondsuit$the method exhibits convergence of order $p > 1$, if

$$\exists C : \text{Error}_P(t+1) \leq C \left(\text{Error}_P(t)\right)^p$$

Convergence of order 2 is called *quadratic*. For example,

$$\text{Error}_P(t) = e^{-ap^t}$$

converges to 0 with order $p$.

Informal explanation: When the method converges, $\text{Error}_P(t)$ goes to 0 as $t \to \infty$, that is, eventually the decimal representation of $\text{Error}_P(t)$ has zero before the decimal dot and more and more zeros after the dot; the number of zeros following the decimal dot is called *the number of accuracy digits in the corresponding approximate solution*. Traditional classification of rates of convergence is based on *how many steps, asymptotically, is required to add a new accuracy digit to the existing ones.*

$\diamondsuit$With *sublinear* convergence, the "price" of accuracy digit grows with the position of the digit. For example, with rate of convergence $O(1/t)$ every new accuracy digit is 10 times more expensive, in terms of # of steps, than its predecessor.

$\diamondsuit$With *linear* convergence, every accuracy digit has the same price, proportional to $\frac{1}{\ln\left(\frac{1}{\text{convergence ratio}}\right)}$. Equivalently: every step of the method adds a fixed number $r$ of accuracy digits (for $q$ not too close to 0, $r \approx 1 - q$);

$\diamondsuit$With *superlinear* convergence, every subsequent accuracy digit eventually becomes cheaper than its predecessor – the price of accuracy digit goes to 0 as the position of the digit grows. Equivalently, every additional step adds more and more accuracy digits.

$\diamondsuit$With convergence of order $p > 1$, the price of accuracy digit not only goes to 0 as the position $k$ of the digit grows, but does it rapidly enough – in a geometric progression. Equivalently, eventually every additional step of the method *multiplies by* $p$ the number of accuracy digits.

♣ With the traditional approach, the convergence properties of a method are the better the higher is the "rank" of the method in the above classification. Given a family of problems, traditionally it is thought that linearly converging on every problem of the family method is faster than a sublinearly converging, superlinearly converging method is faster than a linearly converging one, etc.

♣ Note: Usually we are able to *prove existence* of parameters $C$ and $q$ quantifying linear convergence:

$$\mathrm{Error}_P(t) \leq Cq^t$$

or convergence of order $p > 1$:

$$\mathrm{Error}_P(t+1) \leq C(\mathrm{Error}_P(t))^p,$$

but are unable to find numerical values of these parameters – they may depend on "unobservable" characteristics of a particular problem we are solving. As a result, traditional "quantification" of convergence properties is *qualitative* and *asymptotical*.

## Solvable Case of MP – Convex Programming

♣ We have seen that *as applied to general MP programs*, optimization methods have a number of severe *theoretical* limitations, including the following major ones:
♢Unless exhaustive search (completely unrealistic in high-dimensional optimization) is used, there are no guarantees of approaching *global* solution
♢Quantification of convergence properties is of asymptotical and qualitative character. As a result, the most natural questions like:

> *We should solve problems of such and such structure with such and such sizes and the data varying in such and such ranges. How many steps of method X are sufficient to solve problems within such and such accuracy?*

usually do not admit theoretically valid answers.

♣ In spite of their *theoretical* limitations, *in reality* traditional MP algorithms allow to solve many, if not all, MP problems of real-world origin, including those with many thousands variables and constraints.
♣ Moreover, there exists a "solvable case" when practical efficiency admits solid theoretical guarantees – the case of Convex Programming.

• Here is a typical "Convex Programming" result:

Assume we are solving a Convex Programming program

$$\mathrm{Opt} = \min_x \left\{ f(x) : g_j(x) \le 0,\, j \le m, |x_i| \le 1, i \le n \right\}.$$

where the objective and the constraints are normalized by the requirement

$$|x_i| \le 1,\, i \le n \Rightarrow\ |f(x)| \le 1,\, |g_j(x)| \le 1,\, j \le m$$

Given $\epsilon \in (0,1)$, one can find an $\epsilon$-solution $x^\epsilon$ to the problem:

$$\underbrace{|x_i^\epsilon| \le 1}_{\forall i \le n}\ \&\ \underbrace{g_j(x^\epsilon) \le \epsilon}_{\forall j \le m}\ \&\ f(x^\epsilon) - \mathrm{Opt} < \epsilon$$

in no more than

$$2n^2 \ln\left(\frac{2n}{\epsilon}\right)$$

steps, with a single computation of the values and the first order derivatives of $f, g_1, ..., g_m$ at a point and $100(m+n)n$ additional arithmetic operations per step.

# Line Search

♣ Line Search is a common name for techniques for *one-dimensional* "simply constrained" optimization, specifically, for problems

$$\min_x \{f(x) : a \le x \le b\},$$

where $[a,b]$ is a given segment on the axis (sometimes, we shall allow for $b = +\infty$), and $f$ is a function which is at least once continuously differentiable on $(a,b)$ and is continuous at the segment $[a,b]$ (on the ray $[a,\infty)$, if $b = \infty$).

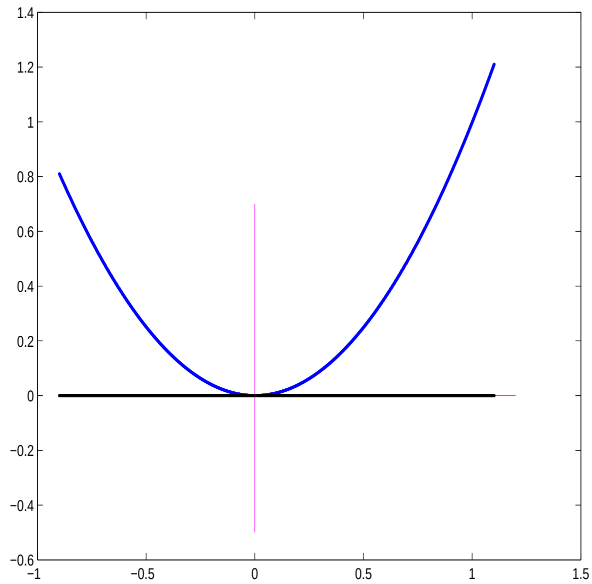
♣ Line search is used, as a subroutine, in many algorithms for multi-dimensional optimization.

$$\min_{a \le x \le b} f(x) \qquad (P)$$

♣ <u>Zero-order line search.</u> In zero-order line search one uses the values of the objective $f$ in $(P)$ and does not use its derivatives.

♠ To ensure well-posedness of the problem, assume that the objective is *unimodal*, that is, possesses a unique local minimizer $x_*$ on $[a, b]$.

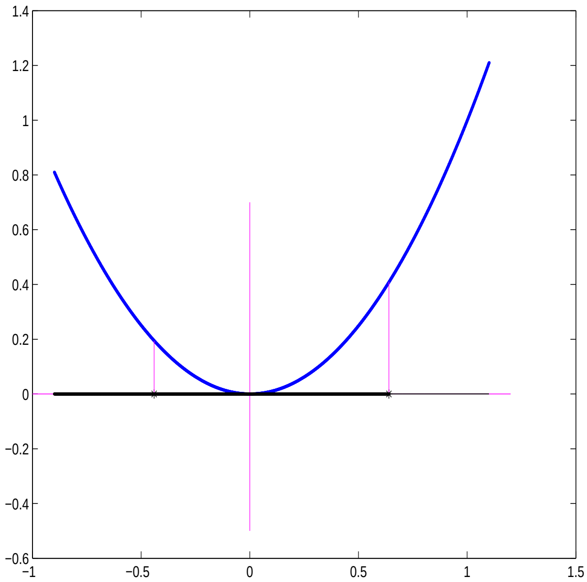
<u>Equivalently:</u> There exists a unique point $x_* \in [a, b]$ such that $f(x)$ strictly decreases on $[a, x_*]$ and strictly increases on $[x_*, b]$:



Main observation: Let $f$ be unimodal on $[a, b]$, and assume we know $f(x')$, $f(x'')$ for certain $x', x''$ with

$$a < x' < x'' < b.$$

♢If $f(x'') \geq f(x')$, then $f(x) > f(x'')$ for $x > x''$, so that the minimizer belongs to $[a, x'']$:




♢Similarly, if $f(x'') < f(x')$, then $f(x > f(x')$ when $x < x'$, so that the minimizer belongs to $[x', b]$.

♠ In both cases, two computations of $f$ at $x'$, $x''$ allow to reduce the initial "search domain" with a smaller one ($[a, x'']$ or $[x', b]$).

♣ Choosing $x'$, $x''$ so that they split $[a_0, b_0] = [a, b]$ into three equal segments, computing $f(x')$, $f(x'')$ and comparing them to each other, we can build a new segment $[a_1, b_1] \subset [a_0, b_0]$ such that

♢the new segment is a *localizer* – it contains the solution $x_*$

♢the length of the new localizer is 2/3 of the length of the initial localizer $[a_0, b_0] = [a, b]$.

♠ On the new localizer, same as on the original one, the objective is unimodal, and we can iterate our construction.

♠ In $N \geq 1$ steps ($2N$ computations of $f$), we shall reduce the size of localizer by factor $(2/3)^N$, that is, we get *linearly converging*, in terms of the argument, algorithm with the convergence ratio

$$q = \sqrt{2/3} = 0.8165...$$

Can we do better ? - YES!

$$\left.\begin{array}{c}[a_{t-1}, b_{t-1}] \\ x_t' < x_t''\end{array}\right\} \Rightarrow f(x_t'), f(x_t'') \Rightarrow \left\{\begin{array}{l}[a_t, b_t] = [a_{t-1}, x_t''] \\ [a_t, b_t] = [x_t', b_{t-1}]\end{array}\right.$$

♣ Observe that one of two points at which we compute $f$ at a step becomes the end-point of the new localizer, while the other one is an interior point of this localizer, *and therefore we can use it as the one of two points where $f$ should be computed at the next step!*
With this approach, only the very first step costs 2 function evaluations, while the subsequent steps cost just 1 evaluation each!
♠ Let us implement the idea in such a way that all search points will divide respective localizers in a fixed proportion:

$$x' - a = b - x'' = \theta(b - a)$$

The proportion is given by the equation

$$\theta \equiv \frac{x' - a}{b - a} = \frac{x'' - x'}{b - x'} \equiv \frac{1 - 2\theta}{1 - \theta} \Rightarrow \theta = \frac{3 - \sqrt{5}}{2}.$$

♣ We have arrived at *golden search*, where the search points $x_{t-1}$, $x_t$ of step $t$ are placed in the current localizer $[a_{t-1}, b_{t-1}]$ according to

$$\frac{x' - a}{b - a} = \frac{b - x''}{b - a} = \frac{3 - \sqrt{5}}{2}$$

In this method, a step reduces the error (the length of localizer) by factor $1 - \frac{3-\sqrt{5}}{2} = \frac{\sqrt{5}-1}{2}$. The convergence ratio is about

$$\frac{\sqrt{5} - 1}{2} \approx 0.6180...$$

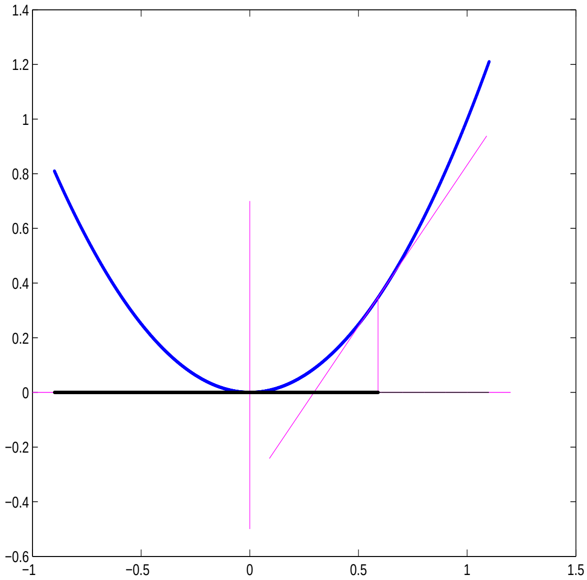
$$\min_x \{f(x) : a \le x \le b\},$$

♣ First order line search: Bisection. Assume that $f$ is differentiable on $(a,b)$ and *strictly unimodal*, that is, it is unimodal, $x_* \in (a,b)$ and $f'(x) < 0$ for $a < x < x_*$, $f'(x) > 0$ for $x_* < x < b$.

Let both $f$ and $f'$ be available. In this case the method of choice in *Bisection*.

♠ Main observation: Given $x_1 \in [a,b] \equiv [a_0, b_0]$, let us compute $f'(x_1)$.

♢If $f'(x_1) > 0$, then, from strict unimodality, $f(x) > f(x_1)$ to the right of $x_1$, thus, $x_*$ belongs to $[a, x_1]$:



♢Similarly, if $f'(x_1) \leq 0$, then $f(x) > f(x_1)$ for $x < x_1$, and $x_*$ belongs to $[a, x_1]$.

♠ In both cases, we can replace the original localizer $[a, b] = [a_0, b_0]$ with a smaller localizer $[a_1, b_1]$ and then iterate the process.

In Bisection, the point $x_t$ where at step $t$ $f'(x_t)$ is computed, is the midpoint of $[a_{t-1}, b_{t-1}]$, so that every step reduces localizer's length by factor 2.

♣ Clearly, Bisection converges linearly in terms of argument with convergence ratio 0.5:

$$a_t - x_* \leq 2^{-t}(b_0 - a_0).$$

## Inexact Line Search

♣ Many algorithms for multi-dimensional minimization which use Line Search as a sub-routine, in the following way:

♢given current iterate $x_t \in \mathbb{R}^n$, the algorithm defines a *search direction* $d_t \in \mathbb{R}^n$ which is a direction of decrease of $f$:

$$d_t^T \nabla f(x_t) < 0.$$

Then Line Search is invoked to minimize the one-dimensional function

$$\phi(s) = f(x_t + \gamma d_t)$$

over $\gamma \geq 0$; the resulting $\gamma = \gamma^t$ defines the stepsize along the direction $d_t$, so that the new iterate of the outer algorithm is

$$x_{t+1} = x_t + \gamma^t d_t.$$

♠ In many situations of this type, there is no necessity in exact minimization in $\gamma$; an "essential" reduction in $\phi$ is sufficient.

♣ Standard way to define (and to achieve) "essential reduction" is given by *Armijo's rule*:

Let $\phi(\gamma)$ be continuously differentiable function of $\gamma \geq 0$ such that $\phi'(0) > 0$, and let $\epsilon \in (0,1)$, $\eta > 1$ be parameters (popular choice is $\epsilon = 0.2$ and $\eta = 2$ or $\eta = 10$).
We say that a stepsize $\gamma > 0$ is *appropriate*, if

$$\phi(\gamma) \leq \phi(0) + \epsilon\gamma\phi'(0), \qquad (*)$$

and is *nearly maximal*, if $\eta$ times larger step is *not* appropriate:

$$\phi(\eta\gamma) > \phi(0) + \epsilon\eta\gamma\phi'(0). \qquad (**)$$

A stepsize $\gamma > 0$ passes Armijo test (reduces $\phi$ "essentially"), if its is both appropriate and nearly maximal.

♠ Fact: Assume that $\phi$ is bounded below on the ray $\gamma > 0$. Then a stepsize passing Armijo rule does exist and can be found efficiently.

♣ Armijo-acceptable step $\gamma > 0$:

$$\phi(\gamma) \leq \phi(0) + \epsilon\gamma\phi'(0) \qquad (*)$$
$$\phi(\eta\gamma) > \phi(0) + \epsilon\eta\gamma\phi'(0) \quad (**)$$

♣ Algorithm for finding Armijo-acceptable step:

**Start:** Choose $\gamma_0 > 0$ and check whether it passes (*). If YES, go to Branch A, otherwise go to Branch B.

**Branch A: $\gamma_0$ satisfies** $(*)$. Testing subsequently the values $\eta\gamma_0$, $\eta^2\gamma_0$, $\eta^3\gamma_0$,... of $\gamma$, stop when the current value for the first time violates $(*)$; the preceding value of $\gamma$ passes the Armijo test.

**Branch B: $\gamma_0$ does not satisfy** $(*)$. Testing subsequently the values $\eta^{-1}\gamma_0$, $\eta^{-2}\gamma_0$, $\eta^{-3}\gamma_0$,... of $\gamma$, stop when the current value for the first time satisfies $(*)$; this value of $\gamma$ passes the Armijo test.

♣ <u>Validation of the algorithm:</u> It is clear that *if the algorithm terminates*, then the result indeed passes the Armijo test. Thus, all we need to verify is that the algorithm eventually terminates.

♢Branch A clearly is finite: here we test the inequality

$$\phi(\gamma) > \phi(0) + \epsilon\gamma\phi'(0)$$

along the sequence $\gamma_i = \eta^i \gamma_0 \to \infty$, and terminate when this inequality is satisfied for the first time. Since $\phi'(0) < 0$ and $\phi$ is below bounded, this indeed will eventually happen.

♢Branch B clearly is finite: here we test the inequality

$$\phi(\gamma) \le \phi(0) + \epsilon\gamma\phi'(0) \qquad (*)$$

along a sequence $\gamma_i = \eta^{-i}\gamma_0 \to +0$ of values of $\gamma$ and terminate when this inequality is satisfied for the first time. Since $\epsilon \in (0,1)$ and $\phi'(0) < 0$, this inequality is satisfied for all small enough positive values of $\gamma$, since

$$\phi(\gamma) = \phi(0) + \gamma\Big[\phi'(0) + \underbrace{R(\gamma)}_{\to 0, \gamma \to +0}\Big].$$

For large $i$, $\gamma_i$ definitely will be "small enough", thus, Branch B is finite.

## Methods for Unconstrained Minimization

♣ Unconstrained minimization problem is

$$f_* = \min_x f(x),$$

where $f$ well-defined and continuously differentiable on the entire $\mathbb{R}^n$.
Note: Most of the constructions to be presented can be straightforwardly extended onto "essentially unconstrained case" where $f$ is continuously differentiable on an open domain $D$ in $\mathbb{R}^n$ and is such that the level sets $\{x \in U : f(x) \le a\}$ are closed.

$$f_* = \min_x f(x) \qquad (P)$$

## Gradient Descent

♣ Gradient Descent is the simplest first order method for unconstrained minimization.
The idea: Let $x$ be a current iterate which is *not* a critical point of $f$: $f'(x) \neq 0$. We have

$$f(x+th) = f(x) + th^T f'(x) + t\|h\|_2 R_x(th)$$
$$[R_x(s) \to 0 \text{ as } s \to 0]$$

Since $f'(x) \neq 0$, the unit antigradient direction $g = -f'(x)/\|f'(x)\|_2$ is a direction of decrease of $f$:

$$\frac{d}{dt}\Big|_{t=0} f(x+tg) = g^T f'(x) = -\|f'(x)\|_2$$

so that shift $x \mapsto x + tg$ along the direction $g$ *locally* decreases $f$ "at the rate" $\|f'(x)\|_2$.
♠ Note: As far as local rate of decrease is concerned, $g$ is the best possible direction of decrease: for any other unit direction $h$, we have

$$\frac{d}{dt}\Big|_{t=0} f(x+th) = h^T f'(x) > -\|f'(x)\|_2.$$

♣ In generic Gradient Descent, we update the current iterate $x$ by a step from $x$ in the antigradient direction which reduces the objective:

$$x_t = x_{t-1} - \gamma_t f'(x_{t-1}),$$

where $\gamma_t$ are positive stepsizes such that

$$f'(x_{t-1}) \neq 0 \Rightarrow f(x_t) < f(x_{t-1}).$$

♠ Standard implementations:

♢Steepest GD:

$$\gamma_t = \operatorname*{argmin}_{\gamma \geq 0} f(x_{t-1} - \gamma f'(x_{t-1}))$$

(slight idealization, except for the case of quadratic $f$)

♢Armijo GD: $\gamma_t > 0$ is such that

$$\begin{array}{l} f(x_{t-1} - \gamma_t f'(x_{t-1}) \leq f(x_{t-1}) - \epsilon\gamma_t \|f'(x_{t-1})\|_2^2 \\ f(x_{t-1} - \eta\gamma_t f'(x_{t-1}) > f(x_{t-1}) - \epsilon\eta\gamma_t \|f'(x_{t-1})\|_2^2 \end{array}$$

(implementable, provided that $f'(x_{t-1}) \neq 0$ and $f(x_{t-1} - \gamma f'(x_{t-1}))$ is below bounded when $\gamma \geq 0$)

Note: By construction, GD is unable to leave a critical point:

$$f'(x_{t-1}) = 0 \Rightarrow x_t = x_{t-1}.$$

♣ Global Convergence Theorem: Assume that the level set of $f$ corresponding to the starting point $x_0$:

$$G = \{x : f(x) \leq f(x_0)\}$$

is compact, and $f$ is continuously differentiable in a neighbourhood of $G$. Then for both SGD and AGD:

♢the trajectory $x_0, x_1, ...$ of the method, started at $x_0$, is well-defined and never leaves $G$ (and thus is bounded);

♢the method is monotone:

$$f(x_0) \geq f(x_1) \geq ...$$

and inequalities are strict, unless method reaches a critical point $x_t$, so that $x_t = x_{t+1} = x_{t+2} = ...$

♢Every limiting point of the trajectory is a critical point of $f$.

**Sketch of the proof: $\mathbf{1}^0$.** If $f'(x_0) = 0$, the method never leaves $x_0$, and the statements are evident. Now assume that $f'(x_0) \neq 0$. Then the function $\phi_0(\gamma) = f(x_0 - \gamma f'(x_0))$ is below bounded, and the set $\{\gamma \geq 0 : \phi_0(\gamma) \leq \phi_0(0)\}$ is compact along with $G$, so that $\phi_0(\gamma)$ achieves its minimum on the ray $\gamma \geq 0$, and $\phi_0'(0) < 0$. It follows that the first step of GD is well-defined and $f(x_1) < f(x_0)$. The set $\{x : f(x) \leq f(x_1)\}$ is a closed subset of $G$ and thus is compact, and we can repeat our reasoning with $x_1$ in the role of $x_0$, etc. We conclude that the trajectory is well-defined, never leaves $G$ and the objective is strictly decreased, unless a critical point is reached.

**$2^0$.** "all limiting points of the trajectory are critical points of $f$":

Fact: *Let $x \in G$ and $f'(x) \neq 0$. Then there exists $\epsilon > 0$ and a neighbourhood $U$ of $x$ such that for every $x' \in U$ the step $x' \to x'_+$ of the method from $x'$ reduces $f$ by at least $\epsilon$.*

Given Fact, let $x$ be a limiting point of $\{x_i\}$; assume that $f'(x) \neq 0$, and let us lead this assumption to contradiction. By Fact, there exists a neighbourhood $U$ of $x$ such that

$$x_i \in U \Rightarrow f(x_{i+1}) \leq f(x_i) - \epsilon.$$

Since the trajectory visits $U$ infinitely many times and the method is monotone, we conclude that $f(x_i) \to -\infty$, $i \to \infty$, which is impossible, since $G$ is compact, so that $f$ is below bounded on $G$.

## Limiting points of Gradient Descent

♣ Under assumptions of Global Convergence Theorem, limiting points of GD exist, and all of them are critical points of $f$. What kind of limiting points could they be?

♢A nondegenerate maximizer of $f$ cannot be a limiting point of GD, unless the method is started at this maximizer.

♢A saddle point of $f$ is "highly unlikely" candidate to the role of a limiting point. Practical experience says that limiting points are local minimizers of $f$.

♢A nondegenerate global minimizer $x_*$ of $f$, if any, as an "attraction point" of GD: when starting close enough to this minimizer, the method converges to $x_*$.

## Rates of convergence

♣ In general, we cannot guarantee more than convergence to the set of critical points of $f$. A natural error measure associated with this set is

$$\delta^2(x) = \|f'(x)\|_2^2.$$

♠ Definition: Let $U$ be an open subset of $\mathbb{R}^n$, $L \geq 0$ and $f$ be a function defined on $U$. We say that $f$ is $\mathrm{C}^{1,1}(L)$ on $U$, if $f$ is continuously differentiable in $U$ with locally Lipschitz continuous, with constant $L$, gradient:

$$[x, y] \in U \Rightarrow \|f'(x) - f'(y)\|_2 \leq L\|x - y\|_2.$$

We say that $f$ is $\mathrm{C}^{1,1}(L)$ on a set $Q \subset \mathbb{R}^n$, if there exists an open set $U \supset Q$ such that $f$ is $\mathrm{C}^{1,1}(L)$ on $U$.

Note: Assume that $f$ is twice continuously differentiable on $U$. Then $f$ is $\mathrm{C}^{1,1}(L)$ on $U$ iff the norm of the Hessian of $f$ does not exceed $L$:

$$\forall (x \in U, d \in \mathbb{R}^n) : |d^T f''(x) d| \leq L\|d\|_2^2.$$

Theorem. In addition to assumptions of Global Convergence Theorem, assume that $f$ is $C^{1,1}(L)$ on $G = \{x : f(x) \leq f(x_0)\}$. Then

♢For SGD, one has

$$\min_{0\leq\tau\leq t} \delta^2(x_\tau) \leq \frac{2[f(x_0) - f_*]L}{t+1}, \; t = 0, 1, 2, ...$$

♢For AGD, one has

$$\min_{0\leq\tau\leq t} \delta^2(x_\tau) \leq \frac{\eta}{2\epsilon(1-\epsilon)} \cdot \frac{[f(x_0) - f_*]L}{t+1}, \; t = 0, 1, 2, ...$$

Lemma. For $x \in G$, $0 \le s \le 2/L$ one has

$$x - sf'(x) \in G \qquad (1)$$

$$f(x - sf'(x)) \le f(x) - \delta^2(x)s + \frac{L\delta^2(x)}{2}s^2, \qquad (2)$$

There is nothing to prove when $g \equiv -f'(x) = 0$. Let $g \ne 0$, $s_* = \max\{s \ge 0 : x + sg \in G\}$, $\delta^2 = \delta^2(x) = g^T g$. The function

$$\phi(s) = f(x - sf'(x)) : [0, s_*] \to \mathbb{R}$$

is continuously differentiable and satisfies

$$\begin{array}{llll} (a) & \phi'(0) = -g^T g \equiv -\delta^2; & (b) & \phi(s_*) = f(x_0) \\ (c) & |\phi'(s) - \phi'(0)| = |g^T[f'(x+sg) - f'(x)]| \le Ls\delta^2 & & \end{array}$$

Therefore $\phi(s) \le \phi(0) - \delta^2 s + \frac{L\delta^2}{2}s^2$, which is (2). Indeed, setting

$$\theta(s) = \phi(s) - [\phi(0) - \delta^2 s + \frac{L\delta^2}{2}s^2],$$

we have

$$\theta(0) = 0, \theta'(s) = \phi'(s) - \phi'(0) - Ls\delta^2 \underbrace{\le}_{\text{by } (c)} 0.$$

By $(*)$ and $(b)$, we have

$$f(x_0) \le \phi(0) - \delta^2 s_* + \frac{L\delta^2}{2}s_*^2 \le f(x_0) - \delta^2 s_* + \frac{L\delta^2}{2}s_*^2$$
$$\Rightarrow s_* \ge 2/L$$

**Lemma $\Rightarrow$ Theorem:** <u>SGD:</u> By Lemma, we have

$$f(x_t) - f(x_{t+1}) = f(x_t) - \min_{\gamma \geq 0} f(x_t - \gamma f'(x_t))$$
$$\geq f(x_t) - \min_{0 \leq s \leq 2/L} \left[ f(x_t) - \delta^2(x_t) s + \frac{L\delta^2(x_t)}{2} s^2 \right]$$
$$= \frac{\delta^2(x_t)}{2L}$$
$$\Rightarrow f(x_0) - f_* \geq \sum_{\tau=0}^{t} [f(x_\tau) - f(x_{\tau+1})] \geq \sum_{\tau=0}^{t} \frac{\delta^2(x_\tau)}{2L}$$
$$\geq (t+1) \min_{0 \leq \tau \leq t} \delta^2(x_\tau)$$
$$\Rightarrow \min_{0 \leq \tau \leq t} \delta^2(x_\tau) \leq \frac{2L(f(x_0) - f_*)}{t+1}$$

<u>AGD:</u> <u>Claim:</u> $\gamma_{t+1} > \frac{2(1-\epsilon)}{L\eta}$. Indeed, otherwise by Lemma

$$\begin{aligned}
& f(x_t - \gamma_t \eta f'(x_t)) \\
& \leq f(x_t) - \gamma_{t+1} \eta \delta^2(x_t) + \frac{L\delta^2(x_t)}{2} \eta^2 \gamma_{t+1}^2 \\
& = f(x_t) - \underbrace{\left[ 1 - \frac{L}{2} \eta \gamma_{t+1} \right]}_{\geq \epsilon} \eta \gamma_{t+1} \delta^2(x_t) \\
& \leq f(x_t) - \epsilon \eta \gamma_{t+1} \delta^2(x_t)
\end{aligned}$$

which is impossible.

We have seen that $\gamma_{t+1} > \frac{2(1-\epsilon)}{L\eta}$. By Armijo rule,

$$f(x_t)-f(x_{t+1}) \geq \epsilon\gamma_{t+1}\delta^2(x_t) \geq \frac{2\epsilon(1-\epsilon)}{L\eta}\delta^2(x_t);$$

the rest of the proof is as for SGD.

♣ Convex case. In addition to assumptions of Global Convergence Theorem, assume that $f$ is convex.

♠ All critical points of a convex function are its global minimizers

$\Rightarrow$ In Convex case, SGD and AGD converge to the set of global minimizers of $f$: $f(x_t) \to f_*$ as $t \to \infty$, and all limiting points of the trajectory are global minimizers of $f$.

♠ In Convex $\mathrm{C}^{1,1}(L)$ case, one can quantify the global rate of convergence in terms of the residual $f(x_t) - f_*$:

Theorem. Assume that the set $G = \{x : f(x) \le f(x_0)\}$ is convex compact, $f$ is convex on $G$ and $\mathrm{C}^{1,1}(L)$ on this set. Consider AGD, and let $\epsilon \ge 0.5$. Then the trajectory of the method converges to a global minimizer $x_*$ of $f$, and

$$f(x_t) - f_* \le \frac{\eta L \|x_0 - x_*\|_2^2}{4(1-\epsilon)t}, \ t = 1, 2, \ldots$$

♣ Definition: Let $M$ be a convex set in $\mathbb{R}^n$ and $0 < \ell \leq L < \infty$. A function $f$ is called strongly convex, with parameters $\ell, L$, on $M$, if
♢$f$ is $\mathrm{C}^{1,1}(L)$ on $M$
♢for $x, y \in M$, one has

$$[x-y]^T[f'(x)-f'(y)] \geq \ell\|x-y\|_2^2. \qquad (*)$$

The ratio $Q_f = L/\ell$ is called condition number of $f$.

♠ Comment: If $f$ is $\mathrm{C}^{1,1}(L)$ on a convex set $M$, then

$$x, y \in M \Rightarrow |f(y)-[f(x)+(y-x)^T f'(x)]| \leq \frac{L}{2}\|x-y\|_2^2.$$

If $f$ satisfies $(*)$ on a convex set $M$, then

$$\forall x, y \in M : f(y) \geq f(x)+(y-x)^T f'(x)+\frac{\ell}{2}\|y-x\|_2^2.$$

In particular, $f$ is convex on $M$.
$\Rightarrow$ A strongly convex, with parameters $\ell, L$, function $f$ on a convex set $M$ satisfies the relation

$$\begin{array}{l}\forall x, y \in M : f(x)+(y-x)^T f'(x)+\frac{\ell}{2}\|y-x\|_2^2 \\ \leq f(y) \leq f(x)+(y-x)^T f'(x)+\frac{L}{2}\|y-x\|_2^2\end{array}$$

Note: Assume that $f$ is twice continuously differentiable in a neighbourhood of a convex set $M$. Then $f$ is $(\ell, L)$-strongly convex on $M$ iff for all $x \in M$ and all $d \in \mathbb{R}^n$ one has

$$\ell \|d\|_2^2 \leq d^T f''(x) d \leq L \|d\|_2^2$$
$$\Updownarrow$$
$$\lambda_{\min}(f''(x)) \geq \ell, \ \lambda_{\max}(f''(x)) \leq L.$$

In particular,

♠ A quadratic function

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

with positive definite symmetric matrix $A$ is strongly convex with the parameters $\ell = \lambda_{\min}(A)$, $L = \lambda_{\max}(A)$ on the entire space.

♣ GD in strongly convex case.

Theorem. In the strongly convex case, AGD exhibits linear global rate of convergence. Specifically, let the set $G = \{x : f(x) \leq f(x_0)\}$ be closed and convex and $f$ be strongly convex, with parameters $\ell, L$, on $Q$. Then
x $\diamondsuit G$ is compact, and the global minimizer $x_*$ of $f$ exists and is unique;
$\diamondsuit$AGD with $\epsilon \geq 1/2$ converges linearly to $x_*$:

$$\|x_t - x_*\|_2 \leq \theta^t \|x_0 - x_*\|_2$$

$$\theta = \sqrt{\frac{Q_f - (2-\epsilon^{-1})(1-\epsilon)\eta^{-1}}{Q_f + (\epsilon^{-1}-1)\eta^{-1}}} = 1 - O(Q_f^{-1}).$$

Besides this,

$$f(x_t) - f_* \leq \theta^{2t} Q_f [f(x_0) - f_*].$$

♣ SGD in Strongly convex quadratic case.

Assume that $f(x) = \frac{1}{2}x^TAx - b^Tx + c$ is a strongly convex quadratic function: $A = A^T \succ 0$. In this case, SGD becomes implementable and is given by the recurrence

$$\begin{array}{rcl} g_t &=& f'(x_t) = Ax_t - b \\ \gamma_{t+1} &=& \frac{g_t^Tg_t}{g_t^TAg_t} \\ x_{t+1} &=& x_t - \gamma_{t+1}g_t \end{array}$$

and guarantees that

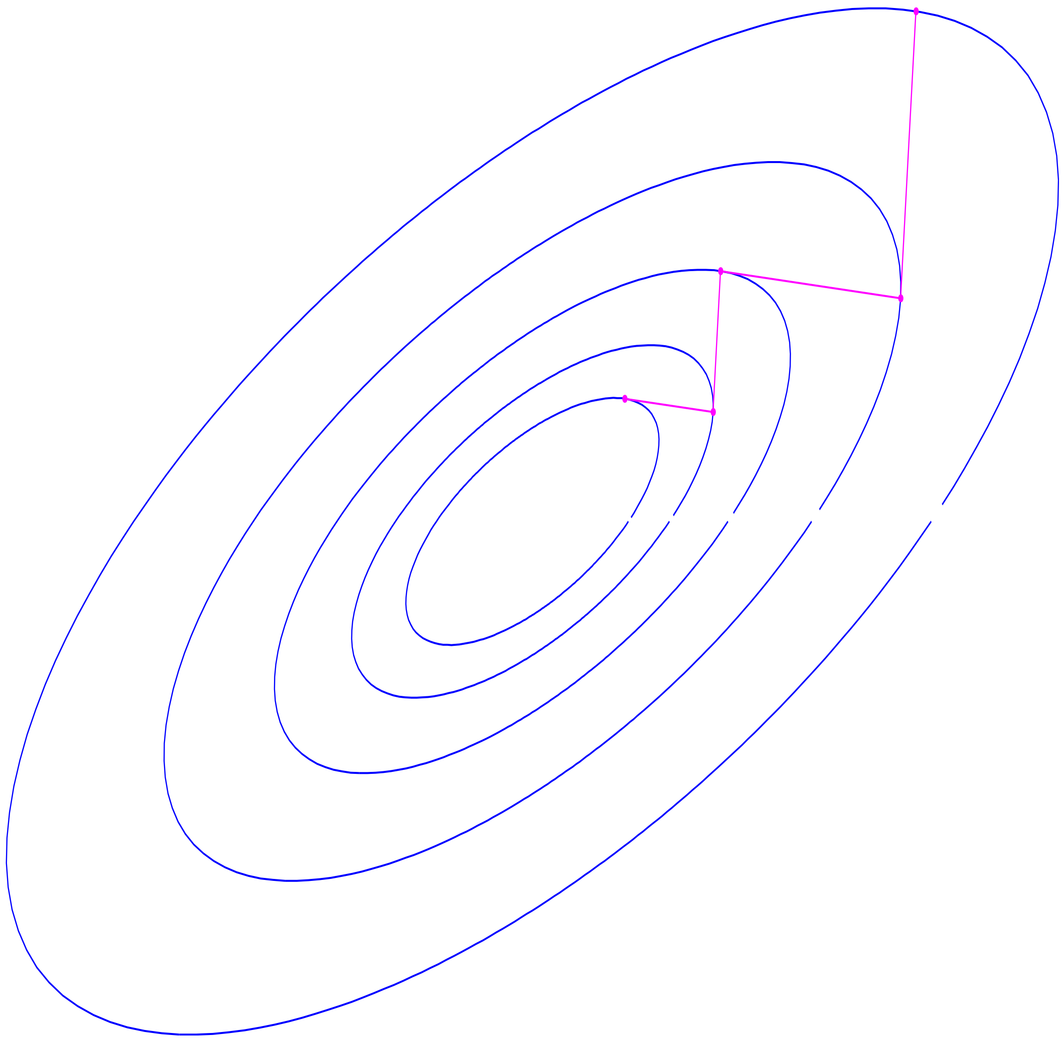
$$\underbrace{f(x_{t+1}) - f_*}_{E_{t+1}} \leq \left[1 - \frac{(g_t^Tg_t)^2}{[g_t^TAg_t][g_t^TA^{-1}g_t]}\right]E_t \leq \left(\frac{Q_f-1}{Q_f+1}\right)E_t$$

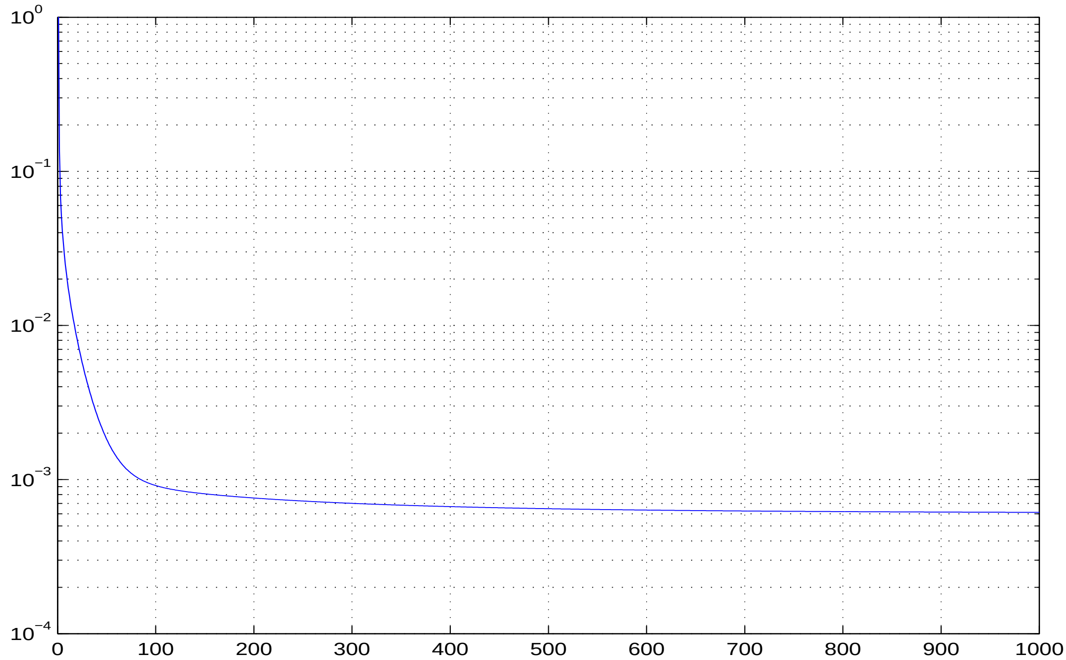
whence

$$f(x_t)-f_* \leq \left(\frac{Q_f-1}{Q_f+1}\right)^{2t}[f(x_0)-f_*],\ t = 1, 2, ...$$

Note: If we know that SGD converges to a *nondegenerate* local minimizer $x_*$ of $f$, then, under mild regularity assumptions, the *asymptotical* behaviour of the method will be as if $f$ were the strongly convex quadratic form

$$f(x) = \text{const} + \frac{1}{2}(x - x_*)^T f''(x_*)(x - x_*).$$






Plot of $\dfrac{f(x_t)-f_*}{(f(x_0)-f_*)\left(\frac{Q_f-1}{Q_f+1}\right)^{2t}}$

SGD as applied to quadratic form with $Q_f = 1000$

$\diamondsuit\, f(x_0) = 2069.4,\ f(x_{999}) = 0.0232$

♣ Summary on Gradient Descent:

♢Under mild regularity and boundedness assumptions, both SGD and AGD converge the set of critical points of the objective.

In the case of $\mathrm{C}^{1,1}(L)$-smooth objective, the methods exhibit non-asymptotical $O(1/t)$-rate of convergence in terms of the error measure $\delta^2(x) = \|f'(x)\|_2^2$.

♢Under the same regularity assumptions, in Convex case the methods converge to the set of glomal minimizers of the objective.

In convex $\mathrm{C}^{1,1}(L)$-case, AGD exhibits non-asymptotical $O(1/t)$ rate of convergence in terms of the residual in the objective $f(x) - f_*$

♢In *Strongly convex case*, AGD exhibits non-asymptotical linear convergence in both the residual in terms of the objective $f(x) - f_*$ and the distance in the argument $\|x - x_*\|_2$. The convergence ratio is $1 - O(1/Q_f)$, where $Q_f$ is the all condition number of the objective. In other words, to get extra accuracy digit, it takes $O(Q_f)$ steps.

♣ <u>Good news on GD:</u>

♠ Simplicity

♠ Reasonable global convergence properties under mild assumptions on the function to be minimized.

♣ Drawbacks of GD:

♠ "Frame-dependence": The method is *not* affine invariant!

♢You are solving the problem $\min_x f(x)$ by GD, starting with $x_0 = 0$, Your first search point will be

$$x_1 = -\gamma_1 f'(0).$$

♢I solve the same problem, but in new variables $y$: $x = Ay$. My problem is $\min_y g(y)$, $g(y) = f(Ax)$, and start with $y_0 = 0$. My first search point will be

$$y_1 = -\widehat{\gamma}_1 g'(0) = -\widehat{\gamma}_1 A^T f'(0).$$

In $x$-variables, my search point will be

$$\widehat{x}_1 = Ay_1 = -\widehat{\gamma}_1 AA^T f'(0)$$

*If $AA^T$ is not proportional to the unit matrix, my search point will, in general, be different from yours!*

♠ "Frame-dependence" is common drawback of nearly all *first order* optimization methods, and this is what makes their rate of convergence, even under the most favourable case of strongly convex objective, sensitive to the condition number of the problem. GD is "hyper-sensitive" to the condition number: When minimizing strongly convex function $f$, the convergence ratio of GD is $1 - O(1/Q_f)$, while for better methods it is $1 - O(1/Q_f^{1/2})$.

## The Newton Method

♣ Consider unconstrained problem

$$\min_x f(x)$$

with *twice* continuously differentiable objective. Assuming second order information available, we approximate $f$ around a current iterate $x$ by the second order Taylor expansion:

$$f(y) \approx f(x) + (y-x)^T f'(x) + \frac{(y-x)^T f''(x)(y-x)}{2}$$

In the Newton method, the new iterate is the minimizer of this quadratic approximation. *If exists*, the minimizer is given by

$$\begin{array}{c}\nabla_y [f(x) + (y-x)^T f'(x) + \frac{(y-x)^T f''(x)(y-x)}{2}] = 0 \\ \Leftrightarrow f''(x)(y-x) = -f'(x) \\ \Leftrightarrow y = x - [f''(x)]^{-1} f'(x)\end{array}$$

We have arrived at the Basic Newton method

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t)$$

(step $t$ is undefined when the matrix $f''(x_t)$ is singular).

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t)$$

♠ Alternative motivation: We seek for a solution to the Fermat equation

$$f'(x) = 0;$$

given current approximate $x_t$ to the solution, we linearize the left hand side around $x_t$, thus arriving at the linearized Fermat equation

$$f'(x_t) + f''(x_t)[x - x_t] = 0$$

and take the solution to this equation, that is, $x_t - [f''(x_t)]^{-1} f'(x_t)$, as our new iterate.

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t) \qquad \text{(Nwt)}$$

Theorem on Local Quadratic Convergence: Let $x_*$ be a nondegenerate local minimizer of $f$, so that $f''(x_*) \succ 0$, and let $f$ be three times continuously differentiable in a neighbourhood of $x_*$. Then the recurrence (Nwt), started close enough to $x_*$, is well-defined and converges to $x_*$ quadratically.

**Proof: 1$^0$.** Let $U$ be a ball centered at $x_*$ where the third derivatives of $f$ are bounded. For $y \in U$ one has

$$\begin{array}{l} \|\nabla f(y) - \nabla^2 f(y)(x_* - y)\|_2 \\ \quad \equiv \|\nabla f(x_*) - \nabla f(y) - \nabla^2 f(y)(x_* - y)\|_2 \\ \quad \leq \beta_1 \|y - x_*\|_2^2 \end{array} \qquad (1)$$

**2$^0$.** Since $f''(x)$ is continuous at $x = x_*$ and $f''(x_*)$ is nonsingular, there exists a ball $U' \subset U$ centered at $x_*$ such that

$$y \in U' \Rightarrow \|[f''(y)]^{-1}\| \leq \beta_2. \qquad (2)$$

<u>Situation:</u> There exists a $r > 0$ and positive constants $\beta_1, \beta_2$ such that

$$\begin{array}{lrcl} & \|y - x_*\| < r & \Rightarrow & \\ (a) & \|\nabla f(y) - \nabla^2 f(y)(x_* - y)\|_2 & \leq & \beta_1 \|y - x_*\|_2^2 \\ (b) & \|[f''(y)]^{-1}\| & \leq & \beta_2 \end{array}$$

**3$^0$.** Let an iterate $x_t$ of the method be close to $x_*$:

$$x_t \in V = \{x : \|x - x_*\|_@ \leq \rho \equiv \min[\frac{1}{2\beta_1\beta_2}, r]\}.$$

We have

$$\begin{array}{l} \|x_{t+1} - x_*\| = \|x_t - x_* - [f''(x_t)]^{-1} f'(x_t)\|_2 \\ = \| \left[ [f''(x_t)]^{-1} [f''(x_t)(x_t - x_*) - f'(x_t)] \right] \|_2 \\ \leq \beta_1\beta_2 \|x_t - x_*\|_2^2 \leq 0.5 \|x_t - x_*\|_2 \end{array}$$

We conclude that the method remains well-defined after step $t$, and converges to $x_*$ quadratically.

♣ A remarkable property of Newton method is affine invariance ("frame independence"): Let $x = Ay + b$ be invertible affine change of variables. Then

$$\begin{array}{rcl} f(x) & \Leftrightarrow & g(y) = f(Ay+b) \\ \bar{x} = Ax + b & \Leftrightarrow & \bar{y} \end{array}$$

$$\begin{array}{l} \bar{y}_+ = \bar{y} - [g''(\bar{y})]^{-1} g'(\bar{y}) \\ = \bar{y} - [A^T f''(\bar{x}) A]^{-1} [A^T f'(\bar{x})] \\ = \bar{y} - A^{-1} [f''(\bar{x})]^{-1} f'(\bar{x}) \\ \Rightarrow A\bar{y}_+ + b = [A\bar{y} + b] - [f''(\bar{x})]^{-1} f'(\bar{x}) \\ = \bar{x} - [f''(\bar{x})]^{-1} f'(\bar{x}) \end{array}$$

♣ <u>Difficulties with Basic Newton method.</u>

The Basic Newton method

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t),$$

*started close enough to nondegenerate local minimizer* $x_*$ *of* $f$, converges to $x_*$ quadratically. However,

♢Even for a nice strongly convex $f$, the method, started not too close to the (unique) local≡global minimizer of $f$, may diverge:

$$f(x) = \sqrt{1+x^2} \Rightarrow x_{t+1} = -x_t^3.$$

$\Rightarrow$ when $|x_0| < 1$, the method converges quadratically (even at order 3) to $x_* = 0$; when $|x_0| > 1$, the method rapidly diverges...

♢When $f$ is not strongly convex, the Newton direction

$$-[f''(x)]^{-1} f'(x)$$

can be undefined or fail to be a direction of decrease of $f$...

♣ As a result of these drawbacks, one needs to modify the Basic Newton method in order to ensure global convergence. Modifications include:

♢Incorporating line search

♢Correcting Newton direction when it is undefined or is not a direction of decrease of $f$.

♣ Incorporating linesearch: Assume that the level set $G = \{x : f(x) \leq f(x_0)\}$ is closed and convex, and $f$ is strongly convex on $G$. Then for $x \in G$ the Newton direction

$$e(x) = -[f''(x)]^{-1} f'(x)$$

is a direction of decrease of $f$, except for the case when $x$ is a critical point (or, which is the same in the strongly convex case, global minimizer) of $f$:

$$\begin{array}{l} f'(x) \neq 0 \Rightarrow \\ \quad e^T(x) f'(x) = -[f'(x)]^T \underbrace{[f''(x)]^{-1}}_{\succ 0} f'(x) < 0. \end{array}$$

In Line Search version of Newton method, one uses $e(x)$ as a search direction rather than the displacement:

$$x_{t+1} = x_t + \gamma_{t+1} e(x_t) = x_t - \gamma_{t+1} [f''(x_t)]^{-1} f'(x_t),$$

where $\gamma_{t+1} > 0$ is the stepsize given by exact minimization of $f$ in the Newton direction or by Armijo linesearch.

Theorem: Let the Level set $G = \{x : f(x) \leq f(x_0)\}$ be convex and compact, and $f$ be strongly convex on $G$. Then Newton method with the Steepest Descent or with the Armijo linesearch converges to the unique global minimizer of $f$.
With proper implementation of the linesearch, convergence is quadratic.

♣ Newton method: Summary

♢Good news: Quadratic asymptotical convergence, provided we manage to bring the trajectory close to a nondegenerate local minimizer

♢Bad news:

— relatively high computational cost, coming from the necessity to compute and to invert the Hessian matrix

— necessity to "cure" the method in the non-strongly-convex case, where the Newton direction can be undefined or fail to be a direction of decrease...

## Modifications of the Newton method

♣ Modifications of the Newton method are aimed at overcoming its shortcomings (difficulties with nonconvex objectives, relatively high computational cost) while preserving its major advantage – rapid asymptotical convergence. There are three major groups of modifications:

♢Modified Newton methods based on second-order information

♢Modifications based on first order information:

— conjugate gradient methods

— quasi-Newton methods

♠ All modifications of Newton method exploit a natural Variable Metric idea.

♣ When speaking about GD, it was mentioned that the method

$$x_{t+1} = x_t - \gamma_{t+1} \underbrace{BB^T}_{A^{-1} \succ 0} f'(x_t) \qquad (*)$$

with nonsingular matrix $B$ has the same "right to exist" as the Gradient Descent

$$x_{t+1} = x_t - \gamma_{t+1} f'(x_t);$$

the former method is nothing but the GD as applied to

$$g(y) = f(By).$$

$$x_{t+1} = x_t - \gamma_{t+1} A^{-1} f'(x_t) \qquad (*)$$

Equivalently: Let $A$ be a positive definite symmetric matrix. We have exactly the same reason to measure the "local directional rate of decrease" of $f$ by the quantity

$$\frac{d^T f'(x)}{\sqrt{d^T d}} \qquad (a)$$

as by the quantity

$$\frac{d^T f'(x)}{\sqrt{d^T A d}} \qquad (b)$$

♢When choosing, as the current search direction, the direction of steepest decrease in terms of $(a)$, we get the anti-gradient direction $-f'(x)$ and arrive at GD.

♢When choosing, as the current search direction, the direction of steepest decrease in terms of $(b)$, we get the "scaled anti-gradient direction" $-A^{-1}f'(x)$ and arrive at "scaled" GD $(*)$.

♣ We have motivated the scaled GD

$$x_{t+1} = x_t - \gamma_{t+1} A^{-1} f'(x_t) \qquad (*)$$

Why not to take one step ahead by consider a generic Variable Metric algorithm

$$x_{t+1} = x_t - \gamma_{t+1} A_{t+1}^{-1} f'(x_t) \qquad \text{(VM)}$$

with "scaling matrix" $A_{t+1} \succ 0$ varying from step to step?

♠ Note: When $A_{t+1} \equiv I$, (VM) becomes the generic Gradient Descent;

When $f$ is strongly convex and $A_{t+1} = f''(x_t)$, (VM) becomes the generic Newton method...

♠ Note: When $x_t$ is *not* a critical point of $f$, the search direction $d_{t+1} = -A_{t+1}^{-1} f'(x_t)$ is a direction of decrease of $f$:

$$d_{t+1}^T f'(x_t) = -[f'(x_t)]^T A_{t+1}^{-1} f'(x_t) < 0.$$

Thus, we have no conceptual difficulties with *monotone* linesearch versions of (VM)...

$$x_{t+1} = x_t - \gamma_{t+1} A_{t+1}^{-1} f'(x_t) \qquad \text{(VM)}$$

♣ It turns out that Variable Metric methods possess good global convergence properties:

Theorem: Let the level set $G = \{x : f(x) \leq f(x_0)\}$ be closed and bounded, and let $f$ be twice continuously differentiable in a neighbourhood of $G$.

Assume, further, that the policy of updating the matrices $A_t$ ensures their *uniform positive definiteness and boundedness*:

$$\exists 0 < \ell \leq L < \infty : \ell I \preceq A_t \preceq LI \; \forall t.$$

Then for both the Steepest Descent and the Armijo versions of (VM) started at $x_0$, the trajectory is well-defined, belongs to $G$ (and thus is bounded), and $f$ strictly decreases along the trajectory unless a critical point of $f$ is reached. Moreover, all limiting points of the trajectory are critical points of $f$.

♣ Implementation via Spectral Decomposition:

♢Given $x_t$, compute $H_t = f''(x_t)$ and then find spectral decomposition of $H_t$:

$$H_t = V_t \mathrm{Diag}\{\lambda_1, ..., \lambda_n\} V_t^T$$

♢Given once for ever chosen tolerance $\delta > 0$, set

$$\widehat{\lambda}_i = \max[\lambda_i, \delta]$$

and

$$A_{t+1} = V_t \mathrm{Diag}\{\widehat{\lambda}_1, ..., \widehat{\lambda}_n\} V_t^T$$

Note: The construction ensures uniform positive definiteness and boundedness of $\{A_t\}_t$, provided the level set $G = \{x : f(x) \leq f(x_0)\}$ is compact and $f$ is twice continuously differentiable in a neighbourhood of $G$.

♣ Levenberg-Marquard implementation:

$$A_{t+1} = \epsilon_t I + H_t,$$

where $\epsilon_t \geq 0$ is chosen to ensure that $A_{t+1} \succeq \delta I$ with once for ever chosen $\delta > 0$.

♢$\epsilon_t$ is found by Bisection as applied to the problem

$$\min \{\epsilon : \epsilon \geq 0, H_t + \epsilon I \succeq \delta I\}$$

♢Bisection requires to check whether the condition

$$H_t + \epsilon I \succ \delta I \Leftrightarrow H_t + (\epsilon - \delta) I \succ 0$$

holds true for a given value of $\epsilon$, and the underlying test comes from Choleski decomposition.

♣ <u>Choleski Decomposition.</u> By Linear Algebra, a symmetric matrix $P$ is $\succ 0$ iff

$$P = DD^T \qquad (*)$$

with lower triangular nonsingular matrix $D$. When Choleski Decomposition $(*)$ exists, it can be found by a simple algorithm as follows:

♢Representation $(*)$ means that

$$p_{ij} = d_i d_j^T,$$

where

$$\begin{array}{rcl} d_i & = & (d_{i1}, d_{i2}, ..., d_{ii}, 0, 0, 0, 0, ..., 0) \\ d_j & = & (d_{j1}, d_{j2}, ..., d_{ji}, ..., d_{jj}, 0, ..., 0) \end{array}$$

are the rows of $D$.

♢In particular, $p_{i1} = d_{11}d_{i1}$, and we can set $d_{11} = \sqrt{p_{11}}$, $d_{i1} = p_{i1}/d_{11}$, thus specifying the first column of $D$.

♢Further, $p_{22} = d_{21}^2 + d_{22}^2$, whence $d_{22} = \sqrt{p_{22} - d_{21}^2}$. After we know $d_{22}$, we can find all remaining entries in the second column of $D$ from the relations

$$p_{i2} = d_{i1}d_{21} + d_{i2}d_{22} \Rightarrow d_{i2} = \frac{p_{i2} - d_{i1}d_{21}}{d_{22}}, \ i > 2.$$

♢We proceed in this way: after the first $(k-1)$ columns in $D$ are found, we fill the $k$-th column according to

$$\begin{aligned} d_{kk} &= \sqrt{p_{kk} - d_{k1}^2 - d_{k2}^2 - \ldots - d_{k,k-1}^2} \\ d_{ik} &= \frac{p_{ik} - d_{i1}d_{k1} - \ldots - d_{i,k-1}d_{k,k-1}}{d_{kk}}, \; i > k. \end{aligned}$$

♠ The outlined process either results in the required $D$, or terminates when you cannot carry out current pivot, that is, when

$$p_{kk} - d_{k1}^2 - d_{k2}^2 - \ldots - d_{k,k-1}^2 \leq 0$$

This "bad termination" indicates that $P$ is not positive definite.

The outlined *Choleski Algorithm* allows to find the Choleski decomposition, if any, in $\approx \frac{n^3}{6}$ a.o. It is used routinely to solve linear systems

$$Px = p \tag{S}$$

with $P \succ 0$. To solve the system, one first computes the Choleski decomposition

$$P = DD^T$$

and then solves $(S)$ by two *back-substitutions*

$$b \mapsto y : Dy = b, \; y \mapsto x : D^T x = y,$$

that is, by solving two triangular systems of equations (which takes just $O(n^2)$ a.o.

Another application of the algorithm (e.g., in Levenberg-Marquardt method) is to check positive definiteness of a symmetric matrix.

Note: The Levenberg-Marquardt method produces uniformly positive definite bounded sequence $\{A_t\}$, provided that the set $G = \{x : f(x) \le f(x_0)\}$ is compact and $f$ is twice continuously differentiable in a neighbourhood of $G$.

♣ The "most practical" implementation of Modified Newton Method is based on running the Choleski decomposition as applied to $H_t = f''(x_t)$. When in course of this process the current pivot (that is, specifying $d_{kk}$) becomes impossible or results in $d_{kk} < \delta$, one increases the corresponding diagonal entry in $H_t$ until the condition $d_{kk} = \delta$ is met.

With this approach, one finds a diagonal correction of $H_t$ which makes the matrix "well positive definite" and ensures uniform positive definiteness and boundedness of the resulting sequence $\{A_t\}$, provided that the set $G = \{x : f(x) \leq f(x_0)\}$ is compact and $f$ is twice continuously differentiable in a neighbourhood of $G$.

## Conjugate Gradient methods

♣ Consider a problem of minimizing a positive definite quadratic form

$$f(x) = \frac{1}{2}x^T Hx - b^T x + c$$

Here is a "conceptual algorithm" for minimizing $f$, or, which is the same, for solving the system

$$Hx = b:$$

*Given starting point $x_0$, let $g_0 = f'(x_0) = Hx_0 - b$, and let*

$$E_k = \mathrm{Lin}\{g_0, Hg_0, H^2 g_0, ..., H^{k-1} g_0\},$$

*and*

$$x_k = \underset{x \in x_0 + E_k}{\mathrm{argmin}}\ f(x).$$

Fact I: Let $k_*$ be the smallest integer $k$ such that $E_{k+1} = E_k$. Then $k_* \leq n$, and $x_{k_*}$ is the unique minimizer of $f$ on $\mathbb{R}^n$

Fact II: One has

$$f(x_k) - \min_x f(x) \leq 4 \left[ \frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1} \right]^{2k} [f(x_0) - \min_x f(x)]$$

<u>Fact III:</u> The trajectory $\{x_k\}$ is given by explicit recurrence:

$\diamondsuit$<u>Initialization:</u> Set

$$d_0 = -g_0 \equiv -f'(x_0) = b - Hx_0;$$

$\diamondsuit$<u>Step $t$:</u> if $g_{t-1} \equiv \nabla f(x_{t-1}) = 0$, terminate, $x_{t-1}$ being the result. Otherwise set

$$\begin{array}{rcl}
\gamma_t & = & -\frac{g_{t-1}^T d_{t-1}}{d_{t-1}^T H d_{t-1}} \\
x_t & = & x_{t-1} + \gamma_t d_{t-1} \\
g_t & = & f'(x_t) \equiv Hx_t - b \\
\beta_t & = & \frac{g_t^T H d_{t-1}}{d_{t-1}^T H d_{t-1}} \\
d_t & = & -g_t + \beta_t d_{t-1}
\end{array}$$

and loop to step $t+1$.

<u>Note:</u> In the above process,

$\diamondsuit$The gradients $g_0, ..., g_{k_*-1}, g_{k_*} = 0$ are mutually orthogonal

$\diamondsuit$The directions $d_0, d_1, ..., d_{k_*-1}$ are $H$-orthogonal:

$$i \neq j \Rightarrow d_i^T H d_j = 0$$

$\diamondsuit$One has

$$\begin{array}{rcl}
\gamma_t & = & \underset{\gamma}{\text{argmin}}\, f(x_{t-1} + \gamma d_{t-1}) \\
\beta_t & = & \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}}
\end{array}$$

♣ Conjugate Gradient method as applied to a strongly convex quadratic form $f$ can be viewed as an iterative algorithm for solving the linear system

$$Hx = b.$$

As compared to "direct solvers", like Choleski Decomposition or Gauss elimination, the advantages of CG are:

♢Ability, in the case of exact arithmetic, to find solution in at most $n$ steps, with a single matrix-vector multiplication and $O(n)$ additional operations per step.

$\Rightarrow$ The cost of finding the solution is at most $O(n)L$, where $L$ is the arithmetic price of matrix-vector multiplication.

Note: When $H$ is sparse, $L << n^2$, and the price of the solution becomes much smaller than the price $O(n^3)$ for the direct LA methods.

♢In principle, there is no necessity to assemble $H$ – all we need is the possibility to multiply by $H$

◇The non-asymptotic error bound

$$f(x_k) - \min_x f(x) \leq 4 \left[ \frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1} \right]^{2k} [f(x_0) - \min_x f(x)]$$

indicates rate of convergence completely independent of the dimension and depending only on the condition number of $H$.

♠ Illustrations:

♢System $1000 \times 1000$, $Q_f = 1.e2$:

| Itr | $f - f_*$ | $\|x - x_*\|_2$ |
|---|---|---|
| 1 | $2.297e+003$ | $2.353e+001$ |
| 11 | $1.707e+001$ | $4.265e+000$ |
| 21 | $3.624e-001$ | $6.167e-001$ |
| 31 | $6.319e-003$ | $8.028e-002$ |
| 41 | $1.150e-004$ | $1.076e-002$ |
| 51 | $2.016e-006$ | $1.434e-003$ |
| 61 | $3.178e-008$ | $1.776e-004$ |
| 71 | $5.946e-010$ | $2.468e-005$ |
| 81 | $9.668e-012$ | $3.096e-006$ |
| 91 | $1.692e-013$ | $4.028e-007$ |
| 94 | $4.507e-014$ | $2.062e-007$ |

$\diamondsuit$System $1000 \times 1000$, $Q_f = 1.e4$:

| Itr | $f - f_*$ | $\|x - x_*\|_2$ |
|---|---|---|
| 1 | $1.471e+005$ | $2.850e+001$ |
| 51 | $1.542e+002$ | $1.048e+001$ |
| 101 | $1.924e+001$ | $4.344e+000$ |
| 151 | $2.267e+000$ | $1.477e+000$ |
| 201 | $2.248e-001$ | $4.658e-001$ |
| 251 | $2.874e-002$ | $1.779e-001$ |
| 301 | $3.480e-003$ | $6.103e-002$ |
| 351 | $4.154e-004$ | $2.054e-002$ |
| 401 | $4.785e-005$ | $6.846e-003$ |
| 451 | $4.863e-006$ | $2.136e-003$ |
| 501 | $4.537e-007$ | $6.413e-004$ |
| 551 | $4.776e-008$ | $2.109e-004$ |
| 601 | $4.954e-009$ | $7.105e-005$ |
| 651 | $5.666e-010$ | $2.420e-005$ |
| 701 | $6.208e-011$ | $8.144e-006$ |
| 751 | $7.162e-012$ | $2.707e-006$ |
| 801 | $7.850e-013$ | $8.901e-007$ |
| 851 | $8.076e-014$ | $2.745e-007$ |
| 901 | $7.436e-015$ | $8.559e-008$ |
| 902 | $7.152e-015$ | $8.412e-008$ |

$\diamondsuit$System $1000 \times 1000$, $Q_f = 1.e6$:

| Itr | $f - f_*$ | $\|x - x_*\|_2$ |
|---|---|---|
| 1 | $9.916e+006$ | $2.849e+001$ |
| 1000 | $7.190e+000$ | $2.683e+000$ |
| 2000 | $4.839e-002$ | $2.207e-001$ |
| 3000 | $4.091e-004$ | $1.999e-002$ |
| 4000 | $2.593e-006$ | $1.602e-003$ |
| 5000 | $1.526e-008$ | $1.160e-004$ |
| 6000 | $1.159e-010$ | $1.102e-005$ |
| 7000 | $6.022e-013$ | $7.883e-007$ |
| 8000 | $3.386e-015$ | $5.595e-008$ |
| 8103 | $1.923e-015$ | $4.236e-008$ |

$\diamondsuit$System $1000 \times 1000$, $Q_f = 1.e12$:

| Itr | $f - f_*$ | $\|x - x_*\|_2$ |
|---|---|---|
| 1 | $5.117e+012$ | $3.078e+001$ |
| 1000 | $1.114e+007$ | $2.223e+001$ |
| 2000 | $2.658e+006$ | $2.056e+001$ |
| 3000 | $1.043e+006$ | $1.964e+001$ |
| 4000 | $5.497e+005$ | $1.899e+001$ |
| 5000 | $3.444e+005$ | $1.851e+001$ |
| 6000 | $2.343e+005$ | $1.808e+001$ |
| 7000 | $1.760e+005$ | $1.775e+001$ |
| 8000 | $1.346e+005$ | $1.741e+001$ |
| 9000 | $1.045e+005$ | $1.709e+001$ |
| 10000 | $8.226e+004$ | $1.679e+001$ |

♣ Non-Quadratic Extensions: CG in the form

$$\begin{array}{rcl} d_0 & = & -g_0 = -f'(x_0) \\ \gamma_t & = & \operatorname*{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\ x_t & = & x_{t-1} + \gamma_t d_{t-1} \\ g_t & = & f'(x_t) \\ \beta_t & = & \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}} \\ d_t & = & -g_t + \beta_t d_{t-1} \end{array}$$

can be applied to *whatever* function $f$, not necessarily quadratic one (Fletcher-Reevs CG), and similarly for another equivalent *in the quadratic case* form:

$$\begin{array}{rcl} d_0 & = & -g_0 = -f'(x_0) \\ \gamma_t & = & \operatorname*{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\ x_t & = & x_{t-1} + \gamma_t d_{t-1} \\ g_t & = & f'(x_t) \\ \beta_t & = & \frac{(g_t - g_{t-1})^T g_t}{g_{t-1}^T g_{t-1}} \\ d_t & = & -g_t + \beta_t d_{t-1} \end{array}$$

(Polak-Ribiere CG).

♠ Being equivalent in the quadratic case, these (and other) forms of CG become different in the non-quadratic case!

♠ Non-quadratic extensions of CG can be used with and without *restarts*.

♢In quadratic case CG, modulo rounding errors, terminate in at most $n$ steps with exact solutions. In non-quadratic case this is not so.

♢In non-quadratic CG with restarts, execution is split in $n$-step *cycles*, and cycle $t+1$ starts from the last iterate $x^t$ of the previous cycle as from the starting point.

In contrast to this, with no restarts the recurrence like

$$\begin{array}{rcl} d_0 &=& -g_0 = -f'(x_0) \\ \gamma_t &=& \operatorname*{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\ x_t &=& x_{t-1} + \gamma_t d_{t-1} \\ g_t &=& f'(x_t) \\ \beta_t &=& \frac{(g_t - g_{t-1})^T g_t}{g_{t-1}^T g_{t-1}} \\ d_t &=& -g_t + \beta_t d_{t-1} \end{array}$$

is never "refreshed".

Theorem: Let the level set $\{x : f(x) \leq f(x_0)\}$ of $f$ be compact and $f$ be twice continuously differentiable in a neighbourhood of $G$. When minimizing $f$ by Fletcher-Reevs or Polak-Ribiere Conjugate Gradients with exact linesearch and restarts,

$\diamondsuit$the trajectory is well-defined and bounded

$\diamondsuit$$f$ never increases

$\diamondsuit$all limiting points of the sequence $x^t$ of concluding iterates of the subsequent cycles are critical points of $f$.

$\diamondsuit$If, in addition, $x^t$ converge to a nondegenerate local minimizer $x_*$ of $f$ and $f$ is 3 times continuously differentiable around $x_*$, then $x^t$ converge to $x_*$ quadratically.

## Quasi-Newton Methods

♣ Quasi-Newton methods are variable metric methods of the generic form

$$x_{t+1} = x_t - \gamma_{t+1} \underbrace{S_{t+1}}_{=A_{t+1}^{-1}} f'(x_t)$$

where $S_{t+1} \succ 0$ and $\gamma_{t+1}$ is given by line-search.

♠ In contrast to Modified Newton methods, in Quasi-Newton algorithms one operates directly on matrix $S_{t+1}$, with the ultimate goal to ensure, under favourable circumstances, that

$$S_{t+1} - [f''(x_t)]^{-1} \to 0,\ t \to \infty. \qquad (*)$$

♠ In order to achieve $(*)$, in Quasi-Newton methods one updates $S_t$ into $S_{t+1}$ in a way which ensures that

♢ $S_{t+1}$ is $\succ 0$

♢ $S_{t+1}(g_t - g_{t-1}) = x_t - x_{t-1}$, where $g_\tau = f'(x_\tau)$.

♣ Generic Quasi-Newton method:

Initialization: Choose somehow starting point $x_0$, matrix $S_1 \succ 0$, compute $g_0 = f'(x_0)$.

Step $t$: given $x_{t-1}$, $g_{t-1} = f'(x_{t-1})$ and $S_t \succ 0$, terminate when $g_{t-1} = 0$, otherwise

♢Set $d_t = -S_t g_{t-1}$ and perform exact line search from $x_{t-1}$ in the direction $d_t$, thus getting new iterate

$$x_t = x_{t-1} + \gamma_t d_t;$$

♢compute $g_t = f'(x_t)$ and set

$$p_t = x_t - x_{t-1},\ q_t = g_t - g_{t-1};$$

♢update $S_t$ into positive definite symmetric matrix $S_{t+1}$ in such a way that

$$S_{t+1} q_t = p_t$$

and loop.

♠ Requirements on the updating rule $S_t \mapsto S_{t+1}$:

♢In order for $d_{t+1}$ to be direction of decrease of $f$, the rule should ensure $S_{t+1} \succeq 0$

♢In the case of strongly convex quadratic $f$, the rule should ensure that $S_t - [f''(\cdot)]^{-1} \to 0$ as $t \to \infty$.

♠ <u>Davidon-Fletcher-Powell method:</u>

$$S_{t+1} = S_t + \frac{1}{p_t^T q_t} p_t p_t^T - \frac{1}{q_t^T S_t q_t} S_t q_t q_t^T S_t.$$

♠ The Davidon-Fletcher-Powell method, as applied to a strongly convex quadratic form, finds exact solution in no more than $n$ steps. The trajectory generated by the method initialized with $S_1 = I$ is exactly the one of the Conjugate Gradient method, so that the DFP (Davidon-Fletcher-Powell) method with the indicated initialization is a Conjugate Gradient method.

♣ The Broyden family.

Broyden-Fletcher-Goldfarb-Shanno updating formula:

$$S_{t+1}^{BFGS} = S_t + \frac{1 + q_t^T S_t q_t}{(p_t^T q_t)^2} p_t p_t^T - \frac{1}{p_t^T q_t} \left[ p_t q_t^T S_t + S_t q_t p_t^T \right]$$

can be combined with the Davidon-Fletcher-Powell formula

$$S_{t+1}^{DFP} = S_t + \frac{1}{q_t^T p_t} p_t p_t^T - \frac{1}{q_t^T S_t q_t} S_t q_t q_t^T S_t.$$

to yield a single-parametric *Broyden* family of updating formulas

$$S_{t+1}^{\phi} = (1 - \phi) S_{t+1}^{DFP} + \phi S_{t+1}^{BFGS}$$

where $\phi \in [0, 1]$ is parameter.

♣ Facts:

♢*As applied to a strongly convex quadratic form $f$, the Broyden method minimizes the form exactly in no more than $n$ steps, $n$ being the dimension of the design vector. If $S_0$ is proportional to the unit matrix, then the trajectory of the method on $f$ is exactly the one of the Conjugate Gradient method.*

♢*all Broyden methods, independently of the choice of the parameter $\phi$, being started from the same pair $(x_0, S_1)$ and equipped with the same exact line search and applied to the same problem, generate the same sequence of iterates (although not the same sequence of matrices $S_t$!).*

♣ *Broyden methods are thought to be the most efficient in paractice versions of the Conjugate Gradient and quasi-Newton methods,* with the pure BFGS method ($\phi = 1$) seemingly being the best.

## Convergence of Quasi-Newton methods

♣ Global convergence of Quasi-Newton methods *without restarts* is proved only for certain versions of the methods and only under strong assumptions on $f$.

• *For methods with restarts*, where the updating formulas are "refreshed" every $m$ steps by setting $S = S_0$, one can easily prove that under our standard assumption that the level set $G = \{x : f(x) \leq f_(x_0)\}$ is compact and $f$ is continuously differentiable in a neighbourhood of $G$, the trajectory of *starting points* of the cycles is bounded, and all its limiting points are critical points of $f$.

♣ Local convergence:

♢*For scheme with restarts,* one can prove that if $m = n$ and $S_0 = I$, then the trajectory of starting points $x^t$ of cycles, if it converges to a nondegenerate local minimizer $x_*$ of $f$ such that $f$ is 3 times continuously differentiable around $x_*$, converges to $x_*$ quadratically.

♢Theorem [Powell, 1976]. Consider the BFGS method without restarts and assume that the method converges to a nondegenerate local minimizer $x^*$ of a three times continuously differentiable function $f$. Then the method converges to $x^*$ superlinearly.

♣ Traditional methods for general constrained problems

$$\min_x \left\{ f(x) : \begin{array}{l} g_j(x) \leq 0,\ j = 1, ..., m \\ h_i(x) = 0,\ i = 1, ..., k \end{array} \right\} \quad (P)$$

can be partitioned into

♢Primal methods, where one mimics unconstrained approach, travelling along the feasible set in a way which ensures progress in objective at every step

♢Penalty/Barrier methods, which reduce constrained minimization to solving a sequence of essentially unconstrained problems

♢Lagrange Multiplier methods, where one focuses on dual problem associated with $(P)$. A posteriori the Lagrange multiplier methods, similarly to the penalty/barrier ones, reduce $(P)$ to a sequence of unconstrained problems, but in a "smart" manner different from the penalty/barrier scheme

♢Sequential Quadratic Programming methods, where one directly solves the KKT system associated with $(P)$ by a kind of Newton method.

## Penalty/Barrier Methods

♣ Penalty Scheme, Equality Constrainsts. Consider equality constrained problem

$$\min_x \{f(x) : h_i(x) = 0,\, i = 1, ..., k\} \qquad (P)$$

and let us "approximate" it by unconstrained problem

$$\min_x f_\rho(x) = f(x) + \underbrace{\frac{\rho}{2}\sum_{i=1}^{k} h_i^2(x)}_{\text{penalty term}} \qquad (P[\rho])$$

$\rho > 0$ is penalty parameter.

Note: (A) On the feasible set, the penalty term vanishes, thus $f_\rho \equiv f$;

(B) When $\rho$ is large and $x$ is infeasible, $f_\rho(x)$ is large:

$$\lim_{\rho\to\infty} f_\rho(x) = \begin{cases} f(x), & x \text{ is feasible} \\ +\infty, & \text{otherwise} \end{cases}$$

$\Rightarrow$ *It is natural to expect that solution of* $(P[\rho])$ *approaches, as* $\rho \to \infty$, *the optimal set of* $(P)$.

♣ Penalty Scheme, General Constraints. In the case of general constrained problem

$$\min_x \left\{ f(x) : \begin{array}{l} h_i(x) = 0,\ i = 1, ..., k \\ g_j \le 0,\ j = 1, ..., m \end{array} \right\}, \qquad (P)$$

the same idea of penalizing the constraint violations results in approximating $(P)$ by unconstrained problem

$$\min_x f_\rho(x) = f(x) + \underbrace{\frac{\rho}{2}\left[\sum_{i=1}^{k} h_i^2(x) + \sum_{j=1}^{m} [g_j(x)^+]^2\right]}_{\substack{\text{penalty}\\ \text{term}}} \qquad (P[\rho])$$

where

$$g_j^+(x) = \max[g_j(x), 0]$$

and $\rho > 0$ is penalty parameter. Here again

$$\lim_{\rho \to \infty} f_\rho(x) = \begin{cases} f(x), & x \text{ is feasible} \\ +\infty, & \text{otherwise} \end{cases}$$

and we again may expect that the solutions of $(P[\rho])$ approach, as $\rho \to \infty$, the optimal set of $(P)$.

♣ Barrier scheme normally is used for *inequality constrained* problems

$$\min_x \left\{ f(x) : g_j(x) \leq 0, j = 1, ..., m \right\} \qquad (P)$$

satisfying "Slater condition": *the feasible set*

$$G = \left\{ x : g_j(x) \leq 0, j \leq m \right\}$$

*of* $(P)$ *possesses a nonempty interior* int$G$ *which is dense in* $G$*, and* $g_j(x) < 0$ *for* $x \in$ int$G$*.*

♠ Given $(P)$, one builds a *barrier* (≡interior penalty) for $G$ – a function $F$ which is well-defined and smooth on int$G$ and blows up to $+\infty$ along every sequence of points $x_i \in \mathrm{int}G$ converging to a boundary point of $G$:

$$x_i \in \mathrm{int}G, \lim_{i\to\infty} x_i = x \notin \mathrm{int}G \Rightarrow F(x_i) \to \infty, i \to \infty.$$

Examples:

♢Log-barrier $F(x) = -\sum_j \ln(-g_j(x))$

♢Carrol Barrier $F(x) = -\sum_j \frac{1}{g_j(x)}$

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\ j = 1, ..., m \right\} \qquad (P)$$

♠ After interior penalty $F$ for the feasible domain of $(P)$ is chosen, the problem is approximated by the "essentially unconstrained" problem

$$\min_{x \in \mathrm{int} G} F_\rho(x) = f(x) + \frac{1}{\rho} F(x) \qquad (P[\rho])$$

When *penalty parameter* $\rho$ is large, the function $F_\rho$ is close to $f$ everywhere in $G$, except for a thin stripe around the boundary.

$\Rightarrow$ It is natural to expect that solutions of $(P[\rho])$ approach the optimal set of $(P)$ as $\rho \to \infty$,

## Investigating Penalty Scheme

♣ Let us focus on equality constrained problem

$$\min_x \{f(x) : h_i(x) = 0,\, i = 1, ..., k\} \qquad (P)$$

and associated penalized problems

$$\min_x f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

(results for general case are similar).

♠ Questions of interest:

♢Whether indeed unconstrained minimizers of the penalized objective $f_\rho$ converge, as $\rho \to \infty$, to the optimal set of $(P)$?

♢What are our possibilities to minimize the penalized objective?

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$
$$\Downarrow$$
$$\min_x f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

Simple fact: Let $(P)$ be feasible, the objective and the constraints in $(P)$ be continuous and let $f$ possess bounded level sets $\{x : f(x) \le a\}$. Let, further $X_*$ be the set of global solutions to $(P)$. Then $X_*$ is nonempty, approximations problems $(P[\rho])$ are solvable, and their global solutions approach $X_*$ as $\rho \to \infty$:

$$\begin{array}{l}\forall \epsilon > 0 \exists \rho(\epsilon) : \rho \ge \rho(\epsilon), x_*(\rho) \ \text{ solves } (P[\rho]) \\ \Rightarrow \text{dist}(x_*(\rho), X_*) \equiv \min\limits_{x_* \in X_*} \|x_*(\rho) - x_*\|_2 \le \epsilon\end{array}$$

**Proof. $1^0$.** By assumption, the feasible set of $(P)$ is nonempty and closed, $f$ is continuous and $f(x) \to \infty$ as $\|x\|_2 \to \infty$. It follows that $f$ attains its minimum on the feasible set, and the set $X_*$ of global minimizers of $f$ on the feasible set is bounded and closed.

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$
$$\Downarrow$$
$$\min_x f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

**$2^0$.** The objective in $(P[\rho])$ is continuous and goes to $+\infty$ as $\|x\|_2 \to \infty$; consequently, $(P[\rho])$ is solvable.

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$
$$\Downarrow$$
$$\min_x f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

**3$^0$.** It remains to prove that, for every $\epsilon > 0$, the solutions of $(P[\rho])$ with large enough value of $\rho$ belong to $\epsilon$-neighbourhood of $X_*$. Assume, on the contrary, that for certain $\epsilon > 0$ there exists a sequence $\rho_i \to \infty$ such that an optimal solution $x_i$ to $(P[\rho_i])$ is at the distance $> \epsilon$ from $X_*$, and let us lead this assumption to contradiction.

$\diamondsuit$Let $f_*$ be the optimal value of $(P)$. We clearly have

$$f(x_i) \le f_{\rho_i}(x_i) \le f_*, \qquad (1)$$

whence $\{x_i\}$ is bounded. Passing to a subsequence, we may assume that $x_i \to \bar{x}$ as $i \to \infty$.

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$
$$\Downarrow$$
$$\min_x f_\rho(x) = f(x) + \tfrac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

$$x_i \in \operatorname*{Argmin}_x f_{\rho_i}(x), x_i \to \bar{x} \notin X_*$$
$$\Rightarrow f(x_i) \le f_{\rho_i}(x_i) \le f_* \qquad (1)$$

♢We claim that $\bar{x} \in X_*$, which gives the desired contradiction. Indeed,
— $\bar{x}$ is feasible, since otherwise

$$\begin{array}{rl} & \lim\limits_{i\to\infty} \underbrace{[f(x_i) + \frac{\rho_i}{2}\|h(x_i)\|_2^2]}_{f_{\rho_i}(x_i)} \\ = & f(\bar{x}) + \lim\limits_{i\to\infty} \frac{\rho_i}{2} \underbrace{\|h(x_i)\|_2^2}_{\to\|h(\bar{x})\|_2^2>0} = +\infty, \end{array}$$

in contradiction to (1);
— $f(\bar{x}) = \lim\limits_{i\to\infty} f(x_i) \le f_*$ by (1); since $\bar{x}$ is feasible for $(P)$, we conclude that $\bar{x} \in X_*$.
♠ Shortcoming of Simple Fact: *In non-convex case, we cannot find/approximate global minimizers of the penalized objective, so that Simple Fact is "unsubstantial"...*

$$\min_x \{f(x) : h_i(x) = 0, \; i = 1, ..., k\} \qquad (P)$$
$$\Downarrow$$
$$\min_x f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

Theorem. Let $x^*$ be a nondegenerate locally optimal solution to $(P)$, i.e., a feasible solution such that

$\diamondsuit$ $f$, $h_i$ are twice continuously differentiable in a neighbourhood of $x_*$,

$\diamondsuit$ the gradients of the constraints taken at $x_*$ are linearly independent,

$\diamondsuit$ at $x_*$, the Second Order Sufficient Optimality condition is satisfied.

There exists a neighbourhood $V$ of $x^*$ and $\bar{\rho} > 0$ such that

$\diamondsuit$ for every $\rho \geq \bar{\rho}$, $f_\rho$ possesses in $V$ exactly one critical point $x_*(\rho)$;

$\diamondsuit$ $x_*(\rho)$ is a nondegenerate local minimizer of $f_\rho$ and a minimizer of $f_\rho$ in $V$;

$\diamondsuit$ $x_*(\rho) \to x_*$ as $\rho \to \infty$.

In addition,

• The local "penalized optimal value"

$$f_\rho(x_*(\rho)) = \min_{x\in V} f_\rho(x)$$

is nondecreasing in $\rho$

Indeed, $f_\rho(\cdot) = f(\cdot) + \frac{\rho}{2}\|h(\cdot)\|_2^2$ grow with $\rho$

• The constraint violation $\|h(x_*(\rho))\|_2$ monotonically goes to 0 as $\rho \to \infty$

Indeed, let $\rho'' > \rho'$, and let $x' = x_*(\rho')$, $x'' = x_*(\rho'')$. Then

$$\begin{array}{l} f(x') + \frac{\rho''}{2}\|h(x')\|_2^2 \geq f(x'') + \frac{\rho''}{2}\|h(x'')\|_2^2 \\ f(x'') + \frac{\rho'}{2}\|h(x'')\|_2^2 \geq f(x') + \frac{\rho'}{2}\|h(x')\|_2^2 \\ \Rightarrow f(x') + f(x'') + \frac{\rho''}{2}\|h(x')\|_2^2 + \frac{\rho'}{2}\|h(x'')\|_2^2 \\ \quad \geq f(x') + f(x'') + \frac{\rho''}{2}\|h(x'')\|_2^2 + \frac{\rho'}{2}\|h(x')\|_2^2 \\ \Rightarrow \frac{\rho''-\rho'}{2}\|h(x')\|_2^2 \geq \frac{\rho''-\rho'}{2}\|h(x'')\|_2^2 \end{array}$$

• The true value of the objective $f(x_*(\rho))$ at $x_*(\rho)$ is nondecreasing in $\rho$

• The quantities $\rho h_i(x_*(\rho))$ converge to optimal Lagrange multipliers of $(P)$ at $x_*$

Indeed,

$$0 = f'_\rho(x_*(\rho)) = f'(x_\rho) + \sum_i (\rho h_i(x_*(\rho)))h'_i(x_*(\rho)).$$

♣ Solving penalized problem

$$\min_x f_\rho(x) \equiv f(x) + \frac{\rho}{2}\|h(x)\|_2^2 \qquad (P[\rho])$$

◊*In principle,* one can solve $(P[\rho])$ by whatever method for unconstrained minimization.

◊However: *The conditioning of $f$ deteriorates as $\rho \to \infty$.*

Indeed, as $\rho \to \infty$, we have

$$\begin{array}{rcl} d^T f''_\rho(\underbrace{x_*(\rho)}_{x})d & = & \underbrace{d^T\left[f''(x) + \sum_i \rho h_i(x) h_i''(x)\right]d}_{\to \nabla_x^2 L(x_*,\mu^*)} \\ & + & \underbrace{\rho \sum_i (d^T h_i'(x))^2}_{\substack{\to \infty, \rho \to \infty \\ \text{except for } d^T h'(x_*) = 0}} \end{array}$$

$\Rightarrow$ slowing down the convergence and/or severe numerical difficulties when working with large penalties...

## Barrier Methods

$$\min_x \{f(x) : x \in G \equiv \{x : g_j(x) \leq 0,\ j = 1, ..., m\}\} \qquad (P)$$
$$\Downarrow$$
$$\min_x F_\rho(x) \equiv f(x) + \frac{1}{\rho} F(x) \qquad (P[\rho])$$

$F$ is interior penalty for $G = \mathrm{cl}(\mathrm{int}G)$:

♢$F$ is smooth on int$G$

♢$F$ tends to $\infty$ along every sequence $x_i \in$ int$G$ converging to a boundary point of $G$.

Theorem. Assume that $G = \mathrm{cl}(\mathrm{int}G)$ is bounded and $f, g_j$ are continuous on $G$. Then the set $X_*$ of optimal solutions to $(P)$ and the set $X_*(\rho)$ of optimal solutions to $(P[\rho])$ are nonempty, and the second set converges to the first one as $\rho \to \infty$: for every $\epsilon > 0$, there exists $\rho = \rho(\epsilon)$ such that

$$\rho \geq \rho(\epsilon), x_*(\rho) \in X_*(\rho) \Rightarrow \mathrm{dist}(x_*(\rho), X_*) \leq \epsilon.$$

♣ In the case of <u>convex</u> program

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

with closed and bounded convex $G$ and convex objective $f$, the domain $G$ can be in many ways equipped with a twice continuously differentiable *strongly convex* penalty $F(x)$.

♠ Assuming $f$ twice continuously differentiable on $\mathrm{int}G$, the aggregate

$$F_\rho(x) = \rho f(x) + F(x)$$

is strongly convex on $\mathrm{int}G$ and therefore attains its minimum at a single point

$$x_*(\rho) = \operatorname*{argmin}_{x \in \mathrm{int}G} F_\rho(x).$$

♠ It is easily seen that the *path* $x_*(\rho)$ is continuously differentiable and converges, as $\rho \to \infty$, to the optimal set of $(P)$.

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

$$\Downarrow$$

$$\min_{x \in \mathrm{int} G} F_\rho(x) = \rho f(x) + F(x) \qquad (P[\rho])$$

$$\Downarrow$$

$$x_*(\rho) = \operatorname*{argmin}_{x \in \mathrm{int} G} F_\rho(x) \underset{\rho \to \infty}{\longrightarrow} \operatorname*{Argmin}_G f$$

♣ in *classical path-following scheme (Fiacco and McCormic, 1967)*, one traces the path $x_*(\rho)$ as $\rho \to \infty$ according to the following generic scheme:

♢Given $(x_i \in \mathrm{int} G, \rho_i > 0)$ with $x_i$ close to $x_*(\rho_i)$,

— update $\rho_i$ into a larger value $\rho_{i+1}$ of the penalty

— minimize $F_{\rho_{i+1}}(\cdot)$, $x_i$ being the starting point, until a new iterate $x_{i+1}$ close to

$$x_*(\rho_{i+1}) = \operatorname*{argmin}_{x \in \mathrm{int} G} F_{\rho_{i+1}}(x)$$

is built, and loop.

♠ To update a tight approximation $x_i$ of $\operatorname{argmin} F_{\rho_i}(x)$ into a tight approximation $x_{i+1}$ of $\operatorname{argmin} F_{\rho_i}(x)$, one can apply to $F_{\rho_{i+1}}(\cdot)$ a method for "essentially unconstrained" minimization, preferably, the Newton method

♠ When Newton method is used, one can try to increase penalty at a "safe" rate, keeping $x_i$ in the domain of quadratic convergence of the Newton method as applied to $F_{\rho_{i+1}}(\cdot)$ and thus making use of fast local convergence of the method.

Questions: • How to choose $F$?

• How to measure closeness to the path?

• *How to ensure "safe" penalty updating without slowing the method down?*

Note: As $\rho \to \infty$, the condition number of $F''_\rho(x_*(\rho))$ may blow up to $\infty$, which, according to the traditional theory of the Newton method, makes the problems of updating $x_i$ into $x_{i+1}$ more and more difficult. Thus, slowing down seems to be unavoidable...

♣ In late 80's, it was discovered that *the classical path-following scheme, associated with properly chosen barriers, admits "safe" implementation without slowing down*. This discovery led to invention of *Polynomial Time Interior Point methods* for convex programs.

♣ Majority of Polynomial Time Interior Point methods heavily exploit the classical path-following scheme; the novelty is in what are the underlying barriers – these are specific *self-concordant* functions especially well suited for Newton minimization.

♣ Let $G$ be a closed convex domain with nonempty interior which does not contain lines. A 3 times continuously differentiable convex function

$$F(x) : \text{int}G \to \mathbb{R}$$

is called self-concordant, if

$\diamondsuit F$ is an interior penalty for $G$:

$$x_i \in \text{int}G, x_i \to x \in \partial G \Rightarrow F(x_i) \to \infty$$

$\diamondsuit F$ satisfies the relation

$$\left|\frac{d^3}{dt^3}\Big|_{t=0} F(x+th)\right| \le 2\left(\frac{d^2}{dt^2}\Big|_{t=0} F(x+th)\right)^{3/2}$$

♠ Let $\vartheta \ge 1$. $F$ is called $\vartheta$-self-concordant barrier for $G$, if, in addition to being self-concordant on $G$, $F$ satisfies the relation

$$\left|\frac{d}{dt}\Big|_{t=0} F(x+th)\right| \le \sqrt{\vartheta}\left(\frac{d^2}{dt^2}\Big|_{t=0} F(x+th)\right)^{1/2}$$

$\vartheta$ is called the *parameter* of s.-c.b. $F$.

♣ Every convex program

$$\min_{x \in G} f(x)$$

can be converted into a convex program with *linear* objective, namely,

$$\min_{t,x} \{t : x \in G, f(x) \le t\}.$$

Assuming that this transformation has been done at the very beginning, we can w.l.o.g. focus on convex program with *linear* objective

$$\min_{x \in G} c^T x \qquad (P)$$

$$\min_{x\in G} c^T x \qquad (P)$$

♣ Assume that $G$ is a closed and bounded convex set with a nonempty interior, and let $F$ be a $\vartheta$-s.c.b. barrier for $G$.

♢Fact I: $F$ is strongly convex on int$G$: $F''(x) \succ 0$ for all $x \in$ int$G$. Consequently,

$$F_\rho(x) \equiv \rho c^T x + F(x)$$

also is strongly convex on int$G$. In particular, the quantity

$$\lambda(x, F_\rho) = \left([F'_\rho(x)]^T \Big[\underbrace{[F''_\rho(x)]}_{=F''(x)}\Big]^{-1} F'_\rho(x)\right)^{1/2}$$

called the Newton decrement of $F_\rho$ at $x$ is well-defined for all $x \in$ int$G$ and all $\rho > 0$.

Note: • $\frac{1}{2}\lambda^2(x, F_\rho) = F_\rho(x) - \min_y \Big[F_\rho(x) + (y-x)^T F'_\rho(x) + \frac{1}{2}(y-x)^T F''_\rho(x)(y-x)\Big]$

• $\lambda(x, F_\rho) \geq 0$ and $\lambda(x, F_\rho) = 0$ iff $x = x_*(\rho)$, so that the Newton decrement can be viewed as a "proximity measure" – a kind of distance from $x$ to $x_*(\rho)$.

$$c_* = \min_{x\in G} c^T x \qquad (P)$$

♣ <u>Fact II:</u> Let $(P)$ be solved via the classical penalty scheme implemented as follows:

♢The barrier underlying the scheme is a $\vartheta$-s.-c.b. $F$ for $G$;

♢"Closeness" of $x$ and $x_*(\rho)$ is specified by the relation $\lambda(x, F_\rho) \leq 0.1$;

♢The penalty update is $\rho_{i+1} = \left(1 + \frac{\gamma}{\sqrt{\vartheta}}\right)\rho_i$, where $\gamma > 0$ is a parameter;

♢To update $x_i$ into $x_{i+1}$, we apply to $F_{\rho_{i+1}}$ the *Damped Newton method* started as $x_i$:

$$x \mapsto x - \frac{1}{1+\lambda(x, F_{\rho_{i+1}})}[F''_{\rho_{i+1}}(x)]^{-1}F'_{\rho_{i+1}}(x)$$

• The method is well-defined, and the number of damped Newton steps in updating $x_i \mapsto x_{i+1}$ depends solely on $\gamma$ (and is as small as 1 for $\gamma = 0.1$)

• One has $c^T x_i - c_* \leq \frac{2\vartheta}{\rho_i}$

$\Rightarrow$ *With the outlined method, it takes $O(\sqrt{\vartheta})$ Newton steps to reduce inaccuracy $c^T x - c_*$ by absolute constant factor!*

♣ <u>Fact III:</u> • Every convex domain $G \subset \mathbb{R}^n$ admits $O(n)$-s.-c.b.

• For typical feasible domains arising in Convex Programming, one can point out explicit "computable" s.-c.b.'s. For example,

♠ Let $G$ be given by $m$ convex quadratic constraints:

$$G = \{x : \underbrace{x^T A_j^T A_j x + 2b_j^T x + c_j}_{g_j(x)} \leq 0,\ 1 \leq j \leq m\}$$

satisfying the Slater condition. Then the logarithmic barrier

$$F(x) = -\sum_{j=1}^{m} \ln(-g_j(x))$$

is $m$-s.-c.b. for $G$.

♠ Let $G$ be given by *Linear Matrix Inequality*

$$G = \{x : \underbrace{A_0 + x_1 A_1 + ... + x_n A_n}_{\mathcal{A}(x): m \times m} \succeq 0\}$$

satisfying the Slater condition: $\mathcal{A}(\bar{x}) \succ 0$ for some $\bar{x}$. Then the log-det barrier

$$F(x) = -\ln \mathrm{Det}(\mathcal{A}(x))$$

is $m$-s.-c.b. for $G$.

♣ Consider an LP

$$\min_z \left\{ c^T z : Az - b \geq 0 \right\} \qquad (P)$$

with $m \times n$ matrix $A$, Null$(A) = \{0\}$, along with the dual problem

$$\max_y \left\{ b^T y : A^T y = c,\ y \geq 0 \right\} \qquad (D)$$

and assume that both problems are strictly feasible:

$$\exists \bar{z} : A\bar{z} - b > 0 \ \&\ \exists y > 0 : A^T y = c$$

<u>Note:</u> Passing from $z$ to "primal slack" $x = Az - b$, we can rewrite $(P)$ as

$$\min_x \left\{ e^T x : x \geq 0, x \in L = \mathrm{Im} A - b \right\} \qquad (P')$$

where $e$ is a vector satisfying $A^T e = c$, so that

$$e^T x = e^T (Az - b) = (A^T e)^T z - \text{const} = c^T z - \text{const}$$

$$\begin{array}{rll} & \min\limits_z \left\{c^Tz : Az - b \geq 0\right\} & (P) \\ \Leftrightarrow & \min\limits_x \left\{e^Tx : x + b \in \mathrm{Im}A, x \geq 0\right\} & (P') \\ & \Updownarrow & \\ & \max\limits_y \left\{b^Ty : \underbrace{A^Ty = c \equiv A^Te}_{\Leftrightarrow y-e\in(\mathrm{Im}A)^\perp}\right\} & (D) \end{array}$$

♠ Let $\Phi(x) = -\sum\limits_{i=1}^{m} \ln x_i$. Equipping the domain of $(P)$ with $m$-s.c.b. $F(z) = \Phi(Az - b)$, consider

$$\begin{array}{rcl} z_*(\rho) & = & \operatorname*{argmin}\limits_z[\rho c^Tz + F(z)] \\ & = & \operatorname*{argmin}\limits_z[\rho e^T(Az - b) + \Phi(Az - b)] \end{array}$$

<u>Observation:</u> The point $x_*(\rho) = Az_*(\rho) - b$ minimizes $\rho e^Tx + \Phi(x)$ over the feasible set of $(P')$:

$$x > 0,\ x + b \in \mathrm{Im}A,\ \rho e + \Phi'(x) \in (\mathrm{Im}A)^\perp.$$

$$\Rightarrow y = y_*(\rho) = -\rho^{-1}\Phi'(x_*(\rho)) \text{ satisfies}$$

$$y > 0,\ y - e \in (\mathrm{Im}A)^\perp,\ -\rho b + \Phi'(y) \in \mathrm{Im}A$$

i.e., the point $y_*(\rho)$ minimizes $-\rho b^Ty + \Phi(y)$ over the feasible set of $(D)$.

♣ We arrive at a nice symmetric picture:

♣ The *primal central path* $x \equiv x_*(\rho)$ which minimizes the *primal aggregate*

$$\rho c^T x + \Phi(x) \qquad [\Phi(x) = -\sum_i \ln x_i]$$

over the primal feasible set is given by

$$x > 0, x + b \in \mathrm{Im} A, \rho c + \Phi'(x) \in (\mathrm{Im} A)^\perp$$

♣ The *dual central path* $y \equiv y_*(\rho)$ which minimizes the *dual aggregate*

$$-\rho b^T y + \Phi(y) \qquad [\Phi(y) = -\sum_i \ln y_i]$$

over the dual feasible set is given by

$$y > 0, y - e \in (\mathrm{Im} A)^\perp, -\rho b + \Phi'(y) \in \mathrm{Im} A$$

♣ The paths are linked by

$$y = -\rho^{-1}\Phi'(x) \Leftrightarrow x = -\Phi'(y_*(\rho)) \Leftrightarrow x_i y_i = \frac{1}{\rho}\,\forall i.$$

$$\Rightarrow \mathrm{DualityGap}(x,y) = x^T y \; = \; \begin{array}{l}[c^T x - \mathrm{Opt}(P)] \\ +[\mathrm{Opt}(D) - b^T y]\end{array}$$

*on the path* is equal to $m\rho^{-1}$.

$$\min_z \left\{c^T z : Az - b \geq 0\right\} \qquad (P)$$

$$\Leftrightarrow \quad \min_x \left\{e^T x : x + b \in \mathrm{Im} A, x \geq 0\right\} \quad (P')$$

$$\Updownarrow$$

$$\max_y \left\{b^T y : \underbrace{A^T y = c \equiv A^T e}_{\equiv y - e \in (\mathrm{Im} A)^\perp}\right\} \qquad (D)$$

♣ Generic Primal-Dual IPM for LP:

♢Given current iterate — primal-dual strictly feasible pair $x^i, y^i$ and value $\rho_i$ of penalty, update it into new iterate $x^{i+1}, y^{i+1}, \rho_{i+1}$ by

♢Updating $\rho_i \mapsto \rho_{i+1} \geq \rho_i$

♢Applying a Newton step to the system

$$x > 0,\ x + b \in \mathrm{Im} A; \quad y > 0,\ y - e \in (\mathrm{Im} A)^\perp$$
$$\mathrm{Diag}\{x\} y = \frac{1}{\rho_+} \underbrace{(1, ..., 1)^T}_{\mathbf{e}}$$

defining the primal-dual central path:

$$x^{i+1} = x^i + \Delta x,\ y^{i+1} = y^i + \Delta y$$

where $\Delta x, \Delta y$ solve the linear system

$$\Delta x \in \mathrm{Im} A,\ \Delta y \in (\mathrm{Im} A)^\perp,$$
$$\mathrm{Diag}\{x^i\} \Delta y + \mathrm{Diag}\{y^i\} \Delta x = \frac{\mathbf{e}}{\rho_{i+1}} - \mathrm{Diag}\{x^i\} y^i$$

$$\min_z \left\{ c^T z : Az - b \geq 0 \right\} \qquad (P)$$

$$\Leftrightarrow \quad \min_x \left\{ e^T x : x + b \in \mathrm{Im} A, x \geq 0 \right\} \quad (P')$$

$$\Updownarrow$$

$$\max_y \left\{ b^T y : \underbrace{A^T y = c \equiv A^T e}_{\equiv y - e \in (\mathrm{Im} A)^\perp} \right\} \qquad (D)$$

♣ The classical path-following scheme as applied to $(P)$ and the $m$-s.c.b. $F(z) = \Phi(Az - b)$ allows to trace the path $z_*(\rho)$ (and thus $x_*(\rho) = Az_*(\rho) - b$).

More advanced *primal-dual* path-following methods *simultaneously* trace the primal and the dual central paths, which results in algorithmic schemes with better practical performance than the one of the "purely primal" scheme.

♣ Both approaches, with proper implementation, result in the best known so far theoretical complexity bounds for LP. According to these bounds, the "arithmetic cost" of generating $\epsilon$-solution to a primal-dual pair of strictly feasible LP's with $m \times n$ matrix $A$ is

$$O(1)mn^2 \ln\left(\frac{mn\Theta}{\epsilon}\right)$$

operations, where $O(1)$ is an absolute constant and $\Theta$ is a data-dependent constant.

♣ In practice, properly implemented primal-dual methods by far outperform the purely primal ones and solve in few tens of Newton iterations real-world LPs with tens and hundreds of thousands of variables and constraints.

## Augmented Lagrangian methods

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$

♣ Shortcoming of penalty scheme: in order to solve $(P)$ to high accuracy, one should work with large values of penalty, which makes the penalized objective

$$f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2$$

difficult to minimize.

♠ Augmented Lagrangian methods use the penalty mechanism in a "smart way", which allows to avoid the necessity to work with very large values of $\rho$.

## Ingredient I: Local Lagrange Duality

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$

♣ Let $x_*$ be a nondegenerate local solution to $(P)$, so that there exists $\lambda$ such that

$$\begin{array}{ll} (a) & \nabla_x L(x_*, \lambda^*) = 0 \\ (b) & d^T \nabla_x^2 L(x_*, \lambda^*) d > 0\ \forall 0 \neq d \in T_{x_*} \end{array}$$
$$\left[ \begin{array}{c} L(x, \lambda) = f(x) + \sum_i \lambda_i h_i(x) \\ T_{x^*} = \{d : d^T h_i'(x) = 0,\ i = 1, ..., k\} \end{array} \right]$$

♠ Assume for the time being that instead of $(b)$, a stronger condition hods true:
(!) *the matrix* $\nabla_x^2 L(x_*, \lambda^*)$ *is positive definite on the entire space*

♣ Under assumption (!), $x_*$ is a nondegenerate unconstrained local minimizer of the smooth function

$$L(\cdot, \lambda^*)$$

and as such can be found by methods for unconstrained minimization.

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1, ..., k\} \qquad (P)$$

♠ Intermediate Summary: If

♢(a) *we are clever enough to guess the vector* $\lambda^*$ *of Lagrange multipliers*,

♢(b) *we are lucky to have* $\nabla_x^2 L(x_*, \lambda^*) \succ 0$,

then $x_*$ can be found by unconstrained optimization technique.

♠ How to become smart when being lucky: Local Lagrange Duality.

Situation: $x_*$ is a nondegenerate local solution to

$$\min_x \{f(x) : h_i(x) = 0, \, i = 1, ..., k\} \qquad (P)$$

and we are lucky:

$$\exists \lambda^* : \nabla_x L(x_*, \lambda^*) = 0, \ \nabla_x^2 L(x_*, \lambda^*) \succ 0 \qquad (!)$$

Fact: Under assumption (!), there exist convex neighbourhood $V$ of $x_*$ and convex neighbourhood $\Lambda$ of $\lambda^*$ such that

(i) For every $\lambda \in \Lambda$, function $L(x, \lambda)$ is strongly convex in $x \in V$ and possesses uniquely defined critical point $x_*(\lambda)$ in $V$ which is continuously differentiable in $\lambda \in \Lambda$. $x_*(\lambda)$ is a nondegenerate local minimizer of $L(\cdot, \lambda)$;

(ii) The function

$$\underline{L}(\lambda) = L(x_*(\lambda), \lambda) = \min_{x \in V} L(x, \lambda)$$

is C$^2$-smooth and concave in $\Lambda$,

$$\underline{L}'(\lambda) = h(x_*(\lambda)),$$

and $\lambda_*$ is a nondegenerate maximizer of $\underline{L}(\lambda)$ on $\Lambda$.

$$\min_x \{f(x) : h_i(x) = 0,\, i = 1, ..., k\} \quad (P)$$
$$\Rightarrow \qquad L(x,\lambda) = f(x) + \sum_i \lambda_i h_i(x)$$

Situation: $\nabla_x L(x_*, \lambda^*) = 0$, $\nabla_x^2 L(x_*, \lambda^*) \succ 0$

$$\begin{array}{rcl} \lambda^* & = & \text{argmax}\, \underline{L}(\lambda) = \min\limits_{x\in V} L(x,\lambda) \\ x_* & = & \text{argmin}_{x\in V} L(x,\lambda) \end{array}$$

$\Rightarrow$ We can solve $(P)$ by maximizing $\underline{L}(\lambda)$ over $\lambda \in \Lambda$ by a first order method for unconstrained minimization.
The first order information on $\underline{L}(\lambda)$ required by the method can be obtained by solving auxiliary unconstrained problems

$$x_*(\lambda) = \underset{x\in V}{\text{argmin}}\, L(x,\lambda)$$

via

$$\begin{array}{rcl} \underline{L}(\lambda) & = & L(x_*(\lambda),\lambda) \\ \underline{L}'(\lambda) & = & h(x_*(\lambda)) \end{array}$$

Note: In this scheme, there are no "large parameters"!
However: How to ensure luck?

♣ How to ensure luck: convexification by penalization

Observe that the problem of interest

$$\min_x \{f(x) : h_i(x) = 0, i = 1, ..., k\} \qquad (P)$$

for every $\rho \geq 0$ is exactly equivalent to

$$\min_x \left\{ f_\rho(x) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 : \begin{array}{c} h_i(x) = 0, \\ i \leq k \end{array} \right\} \qquad (P_\rho)$$

It turns out that

(!) If $x_*$ is a nondegenerate locally optimal solution of $(P)$ and $\rho$ is large enough, then $x_*$ is a locally optimal and "lucky" solution to $(P_\rho)$.

$\Rightarrow$ We can solve $(P)$ by applying the outlined "primal-dual" scheme to $(P_\rho)$, provided that $\rho$ is appropriately large!

Note: Although in our new scheme we do have penalty parameter which should be "large enough", we still have an advantage over the straightforward penalty scheme: in the latter, $\rho$ should go to $\infty$ as $O(1/\epsilon)$ as required inaccuracy $\epsilon$ of solving $(P)$ goes to 0, while in our new scheme a single "large enough" value of $\rho$ will do!

$$\min_x \{f(x) : h_i(x) = 0,\ i = 1,...,k\} \qquad (P)$$

$$\Updownarrow$$

$$\min_x \left\{ f_\rho(x) = f(x) + \tfrac{\rho}{2}\|h(x)\|_2^2 : \begin{array}{c} h_i(x) = 0, \\ i \le k \end{array} \right\} \qquad (P_\rho)$$

Justifying the claim: Let

$$L_\rho(x, \lambda) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 + \sum_i \lambda_i h_i(x)$$

be the Lagrange function of $(P_\rho)$; the Lagrange function of $(P)$ is then $L_o(x, \lambda)$. Given nondegenerate locally optimal solution $x_*$ to $(P)$, let $\lambda^*$ be the corresponding Lagrange multipliers. We have

$$\begin{array}{rcl} \nabla_x L_\rho(x_*, \lambda^*) & = & \nabla_x L_0(x_*, \lambda^*) + \rho \sum\limits_i h_i(x_*) h_i'(x_*) \\ & = & \nabla_x L_0(x_*, \lambda^*) = 0 \\ \nabla_x^2 L_\rho(x_* \lambda^*) & = & \nabla_x^2 L(x_*, \lambda^*) + \rho \sum\limits_i h_i(x_*) h_i''(x_*) \\ & & + \rho \sum\limits_i h_i'(x_*)[h_i'(x_*)]^T \\ & = & \nabla_x^2 L_0(x_*, \rho^*) + \rho H^T H, \end{array}$$

$$H = \begin{bmatrix} [h_1'(x_*)]^T \\ \cdots \\ [h_k'(x_*)]^T \end{bmatrix}$$

$$\nabla_x^2 L_\rho(x_* \lambda^*) = \nabla_x^2 L_0(x_*, \rho^*) + \rho H^T H$$
$$H = \begin{bmatrix} [h_1'(x_*)]^T \\ \cdots \\ [h_k'(x_*)]^T \end{bmatrix}$$

Directions $d$ orthogonal to $h_i'(x_*)$, $i = 1, ..., k$, are exactly the directions $d$ such that $Hd = 0$. Thus,

♢For all $\rho \geq 0$, at $x_*$ the Second Order sufficient optimality condition for $(P_\rho)$ holds true:

$$Hd = 0, d \neq 0 \Rightarrow d^T \nabla_x^2 L_\rho(x_*, \lambda^*) d > 0$$

♢All we need in order to prove that $x^*$ is a "lucky" solution for large $\rho$, is the following Linear Algebra fact:

*Let $Q$ be a symmetric $n \times n$ matrix, and $H$ be a $k \times n$ matrix. Assume that $Q$ is positive definite on the null space of $H$:*

$$d \neq 0, Hd = 0 \Rightarrow d^T Q d > 0.$$

*Then for all large enough values of $\rho$ the matrix $Q + \rho H^T H$ is positive definite.*

*Let $Q$ be a symmetric $n \times n$ matrix, and $H$ be a $k \times n$ matrix. Assume that $Q$ is positive definite on the null space of $H$:*

$$d \neq 0, Hd = 0 \Rightarrow d^T Qd > 0.$$

*Then for all large enough values of $\rho$ the matrix $Q + \rho H^T H$ is positive definite.*

**Proof:** Assume, on the contrary, that there exists a sequence $\rho_i \to \infty$ and $d_i$, $\|d_i\|_2 = 1$:

$$d_i^T [Q + \rho_i H^T H] d_i \leq 0 \ \forall i.$$

Passing to a subsequence, we may assume that $d_i \to d$, $i \to \infty$. Let $d_i = h_i + h_i^\perp$ be the decomposition of $d_i$ into the sum of its projections onto Null($H$) and [Null($H$)]$^\perp$, and similarly $d = h + h^\perp$. Then

$$d_i^T H^T H d_i = \|Hd_i\|_2^2 = \|Hh_i^\perp\|_2^2 \to \|Hh^\perp\|_2^2 \Rightarrow$$
$$0 \geq d_i^T [Q + \rho_i H^T H] d_i = \underbrace{d_i^T Q d_i}_{\to d^T Qd} + \rho_i \underbrace{\|Hh_i^\perp\|_2^2}_{\to \|Hh^\perp\|_2^2} \quad (*)$$

If $h^\perp \neq 0$, then $\|Hh^\perp\|_2 > 0$, and the right hand side in $(*)$ tends to $+\infty$ as $i \to \infty$, which is impossible. Thus, $h^\perp = 0$. But then $0 \neq d \in$ Null($H$) and therefore $d^T Qd > 0$, so that the right hand side in $(*)$ is positive for large $i$, which again is impossible.

## Putting things together: Augmented Lagrangian Scheme

$$\min_x \left\{ f(x) + \frac{\rho}{2}\|h(x)\|_2^2 : \begin{array}{c} h_i(x) = 0 \\ i \le k \end{array} \right\} \qquad (P_\rho)$$

$$\Downarrow$$

$$L_\rho(x,\lambda) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 + \sum_i \lambda_i h_i(x)$$

♣ Generic Augmented Lagrangian Scheme:

*For a given value of $\rho$, solve the dual problem*

$$\max_\lambda \underline{L}_\rho(\lambda) \qquad (D)$$
$$\left[\underline{L}_\rho(\lambda) = \min_x L_\rho(x,\lambda)\right]$$

*by a first order method for unconstrained minimization, getting the first order information for $(D)$ from solving the auxiliary problems*

$$x_\rho(\lambda) = \operatorname*{argmin}_x L_\rho(x,\lambda) \qquad (P^\lambda)$$

*via the relations*

$$\begin{array}{rcl} \underline{L}_\rho(\lambda) & = & L_\rho(x_\rho(\lambda),\lambda) \\ \underline{L}'_\rho(\lambda) & = & h(x_\rho(\lambda)) \end{array}$$

$$\min_x \left\{ f(x) + \frac{\rho}{2}\|h(x)\|_2^2 : \begin{array}{c} h_i(x) = 0 \\ i \le k \end{array} \right\} \qquad (P_\rho)$$

$$\Downarrow$$

$$L_\rho(x,\lambda) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2 + \sum_i \lambda_i h_i(x)$$

$$\Downarrow$$

$$\max_\lambda \left\{ \underline{L}_\rho(\lambda) \equiv \underbrace{\min_x L_\rho(x,\lambda)}_{\text{problem } (P^\lambda)} \right\} \qquad (D)$$

Note: If $\rho$ is large enough and the optimizations in $(P^\lambda)$ and in $(D)$ and are restricted to appropriate convex neighbourhoods of nondegenerate local solution $x_*$ to $(P_\rho)$ and the corresponding vector $\lambda^*$ of Lagrange multipliers, respectively, then

— the objective in $(D)$ is concave and $C^2$, and $\lambda^*$ is a nondegenerate solution to $(D)$

— the objectives in $(P^\lambda)$ are convex and $C^2$, and $x_*(\lambda) = \operatorname*{argmin}_x L_\rho(x,\lambda)$ are nondegenerate local solutions to $(P^\lambda)$

— as the "master method" working on $(D)$ converges to $\lambda^*$, the corresponding primal iterates $x_*(\lambda)$ converge to $x_*$.

♣ <u>Implementation issues:</u>

◇<u>Solving auxiliary problems</u>

$$x_\rho(\lambda) = \operatorname*{argmin}_x L_\rho(x,\lambda) \qquad (P^\lambda)$$

— the best choices are Newton method with linesearch or Modified Newton method, provided that the second order information is available; otherwise, one can use Quasi-Newton methods, Conjugate Gradients, etc.

$\diamondsuit$ Solving the master problem

$$\max_{\lambda}\left\{\underline{L}_\rho(\lambda) \equiv \min_x L_\rho(x,\lambda)\right\} \qquad (D)$$

Surprisingly, the method of choice here is the simplest gradient ascent method with constant step:

$$\lambda^t = \lambda^{t-1} + \rho \underline{L}'_\rho(\lambda^{t-1}) = \lambda^{t-1} + \rho h(x^{t-1}),$$

where $x^{t-1}$ is (approximate) minimizer of $L_\rho(x,\lambda^{t-1})$ in $x$.

Motivation: We have

$$\begin{array}{rcl} 0 & \approx & \nabla_x L_\rho(x^{t-1},\lambda^{t-1}) \\ & = & f'(x^{t-1}) + \sum\limits_i [\lambda_i^{t-1} + \rho h_i(x^{t-1})] h_i'(x^{t-1}) \end{array}$$

which resembles the KKT condition

$$0 = f'(x_*) + \sum_i \lambda_i^* h_i'(x_*).$$

$$\max_{\lambda}\left\{\underline{L}_\rho(\lambda) \equiv \min_x L_\rho(x,\lambda)\right\} \qquad (D)$$

$$\Rightarrow \left\{\begin{array}{rcl} \lambda^t &=& \lambda^{t-1} + \rho h(x^{t-1}) \\ x^{t-1} &=& \mathrm{argmin}_x\, L_\rho(x,\lambda^{t-1}) \end{array}\right. \qquad (*)$$

<u>Justification:</u> Direct computation shows that

$$\begin{array}{rcl} \Psi_\rho &\equiv& \nabla^2_\lambda \underline{L}_\rho(\lambda^*) = -H[Q+\rho H^T H]^{-1}H^T, \\ Q &=& \nabla^2_x L_0(x_*,\lambda^*) \\ H &=& \left[\begin{array}{c} [h_1'(x_*)]^T \\ \cdots \\ [H_k'(x_*)]^T \end{array}\right] \end{array}$$

whence $-\rho\Psi_\rho \to I$ as $\rho\to\infty$.
Consequently, *when $\rho$ is large enough and the starting point $\lambda_0$ in $(*)$ is close enough to $\lambda^*$, $(*)$ ensures linear convergence of $\lambda^t$ to $\lambda^*$ with the ratio tending to 0 as $\rho\to+\infty$.*
Indeed, asymptotically the behaviour of $(*)$ is as if $\underline{L}_\rho(\lambda)$ were quadratic function

$$\mathrm{const} - \frac{1}{2}(\lambda-\lambda^*)^T\Psi_\rho(\lambda-\lambda_*),$$

and for this model recurrence $(*)$ becomes

$$\lambda^t - \lambda^* = \underbrace{(I+\rho\Psi_\rho)}_{\to 0,\rho\to\infty}(\lambda^{t-1}-\lambda^*).$$

♣ <u>Adjusting penalty parameter:</u>

$$\Rightarrow \left\{ \begin{array}{rcl} \lambda^t &=& \lambda^{t-1} + \rho h(x^{t-1}) \\ x^{t-1} &=& \mathrm{argmin}_x\, L_\rho(x, \lambda^{t-1}) \end{array} \right. \qquad (*)$$

When $\rho$ is "large enough", so that $(*)$ converges linearly with reasonable convergence ratio, $\|\underline{L}'_\rho(\lambda^t)\|_2 = \|h(x^t)\|_2$ should go to 0 linearly with essentially the same ratio.

$\Rightarrow$ We can use progress in $\|h(\cdot)\|_2$ to control $\rho$, e.g., as follows: when $\|h(x^t)\|_2 \le 0.25\|h(x^{t-1})\|_2$, we keep the current value of $\rho$ intact, otherwise we increase penalty by factor 10 and recompute $x^t$ with the new value of $\rho$.

## Incorporating Inequality Constraints

♣ Given a general-type constrained problem

$$\min_x \left\{ f(x) : \begin{array}{l} h_i = 0, i \leq m \\ g_j(x) \leq 0, j \leq m \end{array} \right\}$$

we can transform it equivalently into the equality constrained problem

$$\min_{x,s} \left\{ f(x) : \begin{array}{l} h_i(x) = 0, i \leq m \\ g_j(x) + s_j^2 = 0, j \leq k \end{array} \right\}$$

and apply the Augmented Lagrangian scheme to the reformulated problem, thus arriving at Augmented Lagrangian

$$\begin{array}{rcl} L_\rho(x, s; \lambda, \mu) & = & f(x) + \sum\limits_i \lambda_i h_i(x) \\ & & + \sum\limits_j \mu_j [g_j(x) + s_j^2] \\ & & + \frac{\rho}{2} \left[ \sum\limits_i h_i^2(x) + \sum\limits_j [g_j(x) + s_j^2]^2 \right] \end{array}$$

The corresponding dual problem is

$$\max_{\lambda,\mu} \left\{ \underline{L}_\rho(\lambda, \mu) = \min_{x,s} L_\rho(x, s; \mu, \lambda) \right\} \qquad (D)$$

$$
\begin{aligned}
L_\rho(x,s;\lambda,\mu) &= f(x) + \sum_i \lambda_i h_i(x) \\
&\quad + \sum_j \mu_j [g_j(x) + s_j^2] \\
&\quad + \frac{\rho}{2}\left[\sum_i h_i^2(x) + \sum_j [g_j(x) + s_j^2]^2\right]
\end{aligned}
$$

$$\Downarrow$$

$$\max_{\lambda,\mu}\left\{\underline{L}_\rho(\lambda,\mu) \equiv \min_{x,s} L_\rho(x,s;\mu,\lambda)\right\}$$

We can carry out the minimization in $s$ analytically, arriving at

$$
\begin{aligned}
\underline{L}_\rho(\lambda,\mu) &= \min_x \Bigg\{ f(x) + \frac{\rho}{2}\sum_{j=1}^{k}\left(g_j(x) + \frac{\mu_j}{\rho}\right)_+^2 \\
&\quad + \sum_{i=1}^{m} \lambda_i h_i(x) + \frac{\rho}{2}\sum_{i=1}^{m} h_i(x)^2 \Bigg\} \\
&\quad - \sum_{j=1}^{k} \frac{\mu_j^2}{2\rho}
\end{aligned}
$$

where $a_+ = \max[0,a]$.

$\Rightarrow$ *The auxiliary problems arising in the Augmented Lagrangian Scheme are problems in the initial design variables!*

$$\min_x \left\{ f(x) : \begin{array}{l} h_i(x) = 0,\ i \le k \\ g_j(x) \le 0,\ j \le m \end{array} \right\} \qquad (P)$$

$$\Downarrow$$

$$\min_{x,s} \left\{ f(x) : \begin{array}{l} h_i(x) = 0,\ i \le k \\ g_j(x) + s_j^2 = 0,\ j \le m \end{array} \right\} \quad (P')$$

♣ Theoretical analysis of Augmented Lagrangian scheme *for problems with equality constraints* was based on assumption that we are trying to approximate *nondegenerate* local solution. Is it true that when applying reducing the inequality constrained problem to an equality constrained one, we preserve nondegeneracy of the local solution?

Yes!

Theorem. Let $x_*$ be a nondegenerate local solution to $(P)$. Then the point

$$(x_*, s^*) : s_j^* = \sqrt{-g_j(x_*)},\ j = 1, ..., m$$

is a nondegenerate local solution to $(P')$.

## Convex case: Augmented Lagrangians

♣ Consider a convex optimization problem

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\ j = 1, ..., m \right\} \qquad (P)$$

($f$, $g_j$ are convex and $C^2$ on $\mathbb{R}^n$).
Assumption: $(P)$ is solvable and satisfies the Slater condition:

$$\exists \bar{x} : g_j(\bar{x}) < 0 \ j = 1, ..., m$$

♠ In the convex question, the previous local considerations can be globalized due to the Lagrange Duality Theorem.

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\ j = 1, ..., m \right\} \qquad (P)$$

Theorem: Let $(P)$ be convex, solvable and satisfy the Slater condition. Then the dual problem

$$\max_{\lambda \geq 0} \underline{L}(\lambda) \equiv \min_x \underbrace{\left[ f(x) + \sum_j \lambda_j g_j(x) \right]}_{L(x,\lambda)} \qquad (D)$$

possess the following properties:

♢dual objective $\underline{L}$ is concave

♢$(D)$ is solvable

♢for every optimal solution $\lambda^*$ of $(D)$, all optimal solutions of $(P)$ are contained in the set $\mathrm{Argmin}_x\, L(x, \lambda^*)$.

♣ Implications:

♢Sometimes we can build $(D)$ explicitly (e.g., in Linear, Linearly Constrained Quadratic and Geometric Programming). In these cases, we may gain a lot by solving $(D)$ and then recovering solutions to $(P)$ from solution to $(D)$.

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\ j = 1, ..., m \right\} \qquad (P)$$

$$\Downarrow$$

$$\max_{\lambda \geq 0} \underline{L}(\lambda) \equiv \min_x \underbrace{\left[ f(x) + \sum_j \lambda_j g_j(x) \right]}_{L(x,\lambda)} \qquad (D)$$

♢In the general case one can solve $(D)$ numerically by a first order method, thus reducing a problem with *general* convex constraints to one with simple linear constraints. To solve $(D)$ numerically, we should be able to compute the first order information for $\underline{L}$. This can be done via solving the auxiliary problems

$$x_* = x_*(\lambda) = \min_x L(x, \lambda) \qquad (P_\lambda)$$

due to

$$\begin{array}{rcl} \underline{L}(\lambda) & = & L(x_*(\lambda), \lambda) \\ \underline{L}'(\lambda) & = & g(x_*(\lambda)) \end{array}$$

<u>Note:</u> $(P_\lambda)$ is a convex unconstrained program with smooth objective!

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\ j = 1, ..., m \right\} \quad (P)$$

$$\Downarrow$$

$$\max_{\lambda \geq 0} \underline{L}(\lambda) \equiv \min_x \underbrace{\left[ f(x) + \sum_j \lambda_j g_j(x) \right]}_{L(x,\lambda)} \quad (D)$$

♠ Potential difficulties:

♢$\underline{L}(\cdot)$ can be $-\infty$ at some points; how to solve $(D)$?

♢after $\lambda^*$ is found, how to recover optimal solution to $(P)$? In may happen that the set $\operatorname{Argmin}_x L(x, \lambda^*)$ is much wider than the optimal set of $(P)$!

Example: LP. $(P)$ : $\min\limits_x \left\{ c^T x : Ax - b \leq 0 \right\}$. Here

$$\begin{array}{rcl} \underline{L}(\lambda) & = & \min\limits_x \left[ c^T x + (A^T \lambda)^T x - b^T \lambda \right] \\ & = & \begin{cases} -b^T \lambda, & A^T \lambda + c = 0 \\ -\infty, & \text{otherwise} \end{cases} \end{array}$$

— how to solve $(D)$ ???

At the same time, for every $\lambda$ the function $L(x, \lambda)$ is linear in $x$; thus, $\operatorname*{Argmin}\limits_x L(x, \lambda)$ is either $\emptyset$, or $\mathbb{R}^n$ – how to recover $x_*$ ???

♠ Observation: *Both outlined difficulties come from possible non-existence/non-uniqueness of solutions to the auxiliary problems*

$$\min_x L(x,\lambda) \equiv \min_x [f(x) + \sum_j \lambda_j g_j(x)] \quad (P_\lambda)$$

Indeed, if solution $x_*(\lambda)$ to $(P_\lambda)$ exists and is unique and continuous in $\lambda$ on certain set $\Lambda$, then $\underline{L}(\lambda)$ is finite and continuously differentiable on $\Lambda$ due to

$$\begin{array}{rcl} \underline{L}(\lambda) & = & L(x_*(\lambda),\lambda) \\ \underline{L}'(\lambda) & = & g(x_*(\lambda)) \end{array}$$

Besides this, if $\lambda^* \in \Lambda$, then there is no problem with recovering optimal solution to $(P)$ from $\lambda_*$.

Example: Assume that the function

$$r(x) = f(x) + \sum_{j=1}^{k} g_j(x)$$

is locally strongly convex ($r''(x) \succ 0 \ \forall x$) and is such that

$$r(x)/\|x\|_2 \to \infty, \ \ \|x\|_2 \to \infty.$$

Then $x_*(\lambda)$ exists, is unique and is continuous in $\lambda$ on the set $\Lambda = \{\lambda > 0\}$.

♣ In *Augmented Lagrangian* scheme, we ensure local strong convexity of

$$r(\cdot) = f(x) + \text{sum of constraints}$$

by passing from the original problem

$$\min_x \left\{ f(x) : g_j(x) \le 0,\ j = 1, ..., m \right\} \qquad (P)$$

to the *equivalent* problem

$$\min_x \left\{ f(x) : \theta_j(g_j(x)) \le 0,\ j = 1, ..., m \right\} \quad (P')$$

where $\theta_j(\cdot)$ are *increasing strongly convex* smooth functions satisfying the normalization

$$\theta_j(0) = 0,\ \theta_j'(0) = 1.$$

$$\min_x \left\{ f(x) : g_j(x) \le 0,\ j = 1,...,m \right\} \qquad (P)$$
$$\Downarrow$$
$$\min_x \left\{ f(x) : \theta_j(g_j(x)) \le 0,\ j = 1,...,m \right\} \quad (P')$$
$$\left[\theta_j(0) = 0,\ \ \theta_j'(0) = 1\right]$$

Facts:

♢$(P')$ is convex and equivalent to $(P)$

♢optimal Lagrange multipliers for $(P)$ and $(P')$ are the same:

$$\nabla_x[f(x) + \sum_j \lambda_j^* g_j(x)] = 0 \ \&\ \lambda_j^* g_j(x) = 0\,\forall j$$
$$\Updownarrow$$
$$\nabla_x[f(x) + \sum_j \lambda_j^* \theta_j(g_j(x))] = 0 \ \&\ \lambda_j^* g_j(x) = 0\,\forall j$$

♢under mild regularity assumptions,

$$r(x) = f(x) + \sum_j \theta_j(g_j(x))$$

is locally strongly convex and $r(x)/\|x\|_2 \to \infty$ as $\|x\|_2 \to \infty$.

$$\min_x \left\{ f(x) : g_j(x) \leq 0,\, j = 1, ..., m \right\} \qquad (P)$$

$$\Downarrow$$

$$\min_x \left\{ f(x) : \theta_j(g_j(x)) \leq 0,\, j = 1, ..., m \right\} \quad (P')$$

$$\left[\theta_j(0) = 0,\ \theta_j'(0) = 1\right]$$

♣ With the outlined scheme, one passes from the classical Lagrange function of $(P)$

$$L(x, \lambda) = f(x) + \sum_j \lambda_j g_j(x)$$

to the *augmented Lagrange function*

$$\widetilde{L}(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(g_j(x))$$

of the problem, which yields the dual problem

$$\max_{\lambda \geq 0} \underline{\widetilde{L}}(\lambda) \equiv \max_{\lambda \geq 0} \min_x \widetilde{L}(x, \lambda)$$

better suited for numerical solution and recovering a solution to $(P)$ than the usual Lagrange dual of $(P)$.

$$L(x,\lambda) = f(x) + \sum_j \lambda_j g_j(x)$$
$$\Downarrow$$
$$\widetilde{L}(x,\lambda) = f(x) + \sum_j \lambda_j \theta_j(g_j(x))$$
$$\Downarrow$$
$$\max_{\lambda \geq 0} \left[\min_x \widetilde{L}(x,\lambda)\right] \qquad (\widetilde{D})$$

♠ Further flexibility is added by penalty mechanism:

$$\widetilde{L}(x,\lambda) \Rightarrow f(x) + \sum_j \lambda_j \rho^{-1} \theta_j(\rho g_j(x))$$

equivalent to "rescaling"

$$\theta_j(s) \Rightarrow \theta_j^{(\rho)}(s) = \rho^{-1}\theta_j(\rho s).$$

The larger is $\rho$, the faster is convergence of the first order methods as applied to $(\widetilde{D})$ and the more difficult become the auxiliary problems

$$\min_x \left[ f(x) + \sum_j \lambda_j \rho^{-1} \theta_j(\rho g_j(x)) \right]$$

## Sequential Quadratic Programming

♣ SQP is thought of to be the most efficient technique for solving general-type optimization problems with smooth objective and constraints.

♣ SQP methods directly solve the KKT system of the problem by a Newton-type iterative process.

♣ Consider an equality constrained problem

$$\min_x \left\{ f(x) : h(x) = (h_1(x), ..., h_k(x))^T = 0 \right\} \quad (P)$$
$$\Rightarrow L(x, \lambda) = f(x) + h^T(x)\lambda$$

The KKT system of the problem is

$$\begin{array}{rcl} \nabla_x L(x, \lambda) & \equiv & f'(x) + [h'(x)]^T \lambda = 0 \\ \nabla_\lambda L(x, \lambda) & \equiv & h(x) = 0 \end{array} \qquad \text{(KKT)}$$

Every locally optimal solution $x_*$ of $(P)$ which is regular (that is, the gradients $\{h_i'(x_*)\}_{i=1}^k$ are linearly independent) can be extended by properly chosen $\lambda = \lambda^*$ to a solution of (KKT).

♠ (KKT) is a system of nonlinear equations with $n + k$ equations and $n + k$ unknowns. We can try to solve this system by Newton method.

## Newton method for solving nonlinear systems of equations

♣ To solve a system of $N$ nonlinear equations with $N$ unknowns

$$P(u) \equiv (p_1(u), ..., p_N(u))^T = 0,$$

with $C^1$ real-valued functions $p_i$, we act as follows:

*Given current iterate $\bar{u}$, we linearize the system at the iterate, thus arriving at the linearized system*

$$\begin{array}{l} P(\bar{u}) + P'(\bar{u})(u - \bar{u}) \\ \equiv \left[ \begin{array}{c} p_1(\bar{u}) + [p_1'(\bar{u})]^T (u - \bar{u}) \\ \vdots \\ p_N(\bar{u}) + [p_N'(\bar{u})]^T (u - \bar{u}) \end{array} \right] = 0. \end{array}$$

*Assuming the $N \times N$ matrix $P'(\bar{u})$ nonsingular, we solve the linearized system, thus getting the new iterate*

$$\bar{u}^+ = \bar{u} \underbrace{-[P'(\bar{u})]^{-1} P(\bar{u})}_{\substack{\text{Newton} \\ \text{displacement}}};$$

$$\bar{u} \mapsto \bar{u}^+ = \bar{u} - [P'(\bar{u})]^{-1} P(\bar{u}) \qquad (N)$$

Note: The Basic Newton method for unconstrained minimization is nothing but the outlined process as applied to the Fermat equation

$$P(x) \equiv \nabla f(x) = 0.$$

♣ Same as in the optimization case, the Newton method possesses fast local convergence:
Theorem. Let $u_* \in \mathbb{R}^N$ be a solution to the square system of nonlinear equations

$$P(u) = 0$$

with components of $P$ being C$^1$ in a neighbourhood of $u_*$. Assuming that $u_*$ is nondegenerate (i.e., Det$(P'(u_*)) \neq 0$), the Newton method $(N)$, started close enough to $u_*$, converges to $u_*$ superlinearly.
If, in addition, the components of $P$ are C$^2$ in a neighbourhood of $u_*$, then the above convergence is quadratic.

♣ Applying the outlined scheme to the KKT system

$$\begin{array}{rcl}\nabla_x L(x,\lambda) & \equiv & f'(x) + [h'(x)]^T\lambda = 0 \\ \nabla_\lambda L(x,\lambda) & \equiv & h(x) = 0\end{array} \qquad \text{(KKT)}$$

we should answer first of all the following crucial question:

(?) When a KKT point $(x_*, \lambda^*)$ is a nondegenerate solution to (KKT)?

Let us set

$$\begin{array}{l}P(x,\lambda) = \nabla_{x,\lambda} L(x,\lambda) \\ = \left[\begin{array}{c}\nabla_x L(x,\lambda) \equiv f'(x) + [h'(x)]^T\lambda \\ \nabla_\lambda L(x,\lambda) \equiv h(x)\end{array}\right]\end{array}$$

Note that

$$P'(x,\lambda) = \left[\begin{array}{cc}\nabla_x^2 L(x,\lambda) & [h'(x)]^T \\ h'(x) & 0\end{array}\right]$$

$$
\begin{array}{rl}
& \min\limits_x \left\{ f(x) : h(x) = (h_1(x), ..., h_k(x))^T = 0 \right\} \quad (P) \\
\Rightarrow & L(x,\lambda) = f(x) + h^T(x)\lambda \\
\Rightarrow & P(x,\lambda) = \nabla_{x,\lambda} L(x,\lambda) \\
& = \left[ \begin{array}{c} \nabla_x L(x,\lambda) \equiv f'(x) + [h'(x)]^T \lambda \\ \nabla_\lambda L(x,\lambda) \equiv h(x) \end{array} \right] \\
\Rightarrow & P'(x,\lambda) = \left[ \begin{array}{cc} \nabla_x^2 L(x,\lambda) & [h'(x)]^T \\ h'(x) & 0 \end{array} \right]
\end{array}
$$

Theorem. Let $x_*$ be a nondegenerate local solution to $(P)$ and $\lambda^*$ be the corresponding vector of Lagrange multipliers. Then $(x_*, \lambda^*)$ is a nondegenerate solution to the KKT system

$$
P(x,\lambda) = 0,
$$

that is, the matrix $P' \equiv P'(x_*, \lambda^*)$ is nonsingular.

**Proof.** Setting $Q = \nabla_x^2 L(x_*, \lambda^*)$, $H = \nabla h(x_*)$, we have

$$
P' = \left[ \begin{array}{cc} Q & H^T \\ H & 0 \end{array} \right]
$$

$Q = \nabla_x^2 L(x_*, \lambda^*)$, $H = \nabla h(x_*)$,

$$P' = \begin{bmatrix} Q & H^T \\ H & 0 \end{bmatrix}$$

We know that $d \neq 0, Hd = 0 \Rightarrow d^T Q d > 0$ and that the rows of $H$ are linearly independent. We should prove that if

$$0 = P' \begin{bmatrix} d \\ g \end{bmatrix} \equiv \begin{bmatrix} Qd + H^T g \\ Hd \end{bmatrix},$$

then $d = 0$, $g = 0$. We have $Hd = 0$ and

$$0 = Qd + H^T g \Rightarrow d^T Q d + (Hd)^T g = d^T Q d,$$

which, as we know, is possible iff $d = 0$. We now have $H^T g = Qd + H^T g = 0$; since the rows of $H$ are linearly independent, it follows that $g = 0$.

## Structure and interpretation of the Newton displacement

♣ In our case the Newton system

$$P'(u)\Delta = -P(u) \qquad\qquad [\Delta = u^+ - u]$$

becomes

$$\begin{array}{rcl} [\nabla^2_x L(\bar{x},\bar{\lambda})]\Delta x + [\nabla h(\bar{x})]^T \Delta\lambda & = & -f'(\bar{x}) \\ & & -[h'(\bar{x})]^T \lambda \ , \\ {[h'(\bar{x})]\Delta\lambda} & = & -h(\bar{x}) \end{array}$$

where $(\bar{x},\bar{\lambda})$ is the current iterate.
Passing to the variables $\Delta x$, $\lambda^+ = \bar{\lambda} + \Delta\lambda$, the system becomes

$$\begin{array}{rcl} [\nabla^2_x L(\bar{x},\bar{\lambda})]\Delta x + [h'(\bar{x})]^T \lambda^+ & = & -f'(\bar{x}) \\ h'(\bar{x})\Delta x & = & -h(\bar{x}) \end{array}$$

$$
\begin{array}{rcl}
[\nabla_x^2 L(\bar{x},\bar{\lambda})]\Delta x + [h'(\bar{x})]^T\lambda^+ & = & -f'(\bar{x}) \\
h'(\bar{x})\Delta x & = & -h(\bar{x})
\end{array}
$$

Interpretation.

♣ Assume for a moment that we know the optimal Lagrange multipliers $\lambda^*$ and the tangent plane $T$ to the feasible surface at $x_*$. Since $\nabla_x^2 L(x_*,\lambda^*)$ is positive definite on $T$, and $\nabla_x L(x_*,\lambda^*)$ is orthogonal to $T$, $x_*$ is a nondegenerate local minimizer of $L(x,\lambda^*)$ over $x\in T$, and we could find $x_*$ by applying the Newton minimization method to the function $L(x,\lambda^*)$ restricted onto $T$:

$$
\bar{x}\in T \mapsto \bar{x} + \underset{\substack{\Delta x:\\ \bar{x}+\Delta x\in T}}{\text{argmin}} \Big[ L(\bar{x},\lambda^*) + \Delta x^T \nabla_x L(\bar{x},\lambda^*) + \tfrac{1}{2}\Delta x^T \nabla_x^2 L(\bar{x},\lambda^*)\Delta x \Big]
$$

♣ In reality we do not know neither $\lambda^*$, nor $T$, only current approximations $\bar{x}$, $\bar{\lambda}$ of $x_*$ and $\lambda^*$. We can use these approximations to *approximate* the outlined scheme:

• Given $\bar{x}$, we approximate $T$ by the plane

$$\bar{T} = \{y = \bar{x} + \Delta x : [h'(\bar{x})]\Delta x + h(\bar{x}) = 0\}$$

• We apply the outlined step with $\lambda^*$ replaced with $\bar{\lambda}$ and $T$ replaced with $\bar{T}$:

$$\bar{x} \in T \mapsto \bar{x} + \underset{\substack{\Delta x: \\ \bar{x}+\Delta x \in \bar{T}}}{\operatorname{argmin}} \Big[ L(\bar{x}, \bar{\lambda}) + \Delta x^T \nabla_x L(\bar{x}, \bar{\lambda}) \\ + \tfrac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) \Delta x \Big]$$

<u>Note:</u> Step can be simplified to

$$\bar{x} \in T \mapsto \bar{x} \\ + \underset{\substack{\Delta x: \\ \bar{x}+\Delta x \in \bar{T}}}{\operatorname{argmin}} \left[ f(\bar{x}) + \Delta x^T f'(\bar{x}) + \tfrac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) \Delta x \right]$$

due to the fact that for $\bar{x} + \Delta x \in \bar{T}$ one has

$$\begin{aligned} \Delta x^T \nabla_x L(\bar{x}, \bar{\lambda}) &= \Delta x^T f'(\bar{x}) + \bar{\lambda}^T [h'(\bar{x})] \Delta x \\ &= \Delta x^T f'(\bar{x}) - \bar{\lambda}^T h(\bar{x}) \end{aligned}$$

♣ We have arrived at the following scheme: Given approximations $(\bar{x}, \bar{\lambda})$ to a nondegenerate KKT point $x_*, \lambda^*$ of equality constrained problem

$$\min_x \left\{ f(x) : h(x) \equiv (h_1(x), ..., h_k(x))^T = 0 \right\} \qquad (P)$$

solve the auxiliary quadratic program

$$\min_{\Delta x} \left\{ f(\bar{x}) + \Delta x^T f'(\bar{x}) + \frac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) \Delta x : h(\bar{x}) + h'(\bar{x}) \Delta x = 0 \right\} \qquad \text{(QP)}$$

and replace $\bar{x}$ with $\bar{x} + \Delta x_*$.

Note: (QP) is a nice Linear Algebra problem, provided that $\nabla^2 L(\bar{x}, \bar{\lambda})$ is positive definite on the feasible plane $\bar{T} = \{\Delta x : h(\bar{x}) + h'(\bar{x}) \Delta x = 0\}$ (which indeed is the case when $(\bar{x}, \bar{\lambda})$ is close enough to $(x_* \lambda^*)$).

$$\min_x \left\{ f(x) : h(x) \equiv (h_1(x), ..., h_k(x))^T = 0 \right\} \qquad (P)$$

♣ Step of the Newton method as applied to the KKT system of $(P)$:

$$\begin{array}{c}
(\bar{x}, \bar{\lambda}) \mapsto (\bar{x}^+ = \bar{x} + \Delta x, \lambda^+) : \\
\begin{array}{rcl}
[\nabla_x^2 L(\bar{x}, \bar{\lambda})]\Delta x + [h'(\bar{x})]^T \lambda^+ & = & -f'(\bar{x}) \\
h'(\bar{x})\Delta x & = & -h(\bar{x})
\end{array}
\end{array} \qquad (N)$$

♣ Associated quadratic program:

$$\min_{\Delta x} \left\{ f(\bar{x}) + \Delta x^T f'(\bar{x}) + \frac{1}{2}\Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda})\Delta x : h(\bar{x}) + h'(\bar{x})\Delta x = 0 \right\} \qquad (\mathrm{QP})$$

Crucial observation: Let the Newton system underlying $(N)$ be a system with nonsingular matrix. Then the Newton displacement $\Delta x$ given by $(N)$ is the unique KKT point of the quadratic program (QP), and $\lambda^+$ is the corresponding vector of Lagrange multipliers.

$$\begin{array}{rcl} [\nabla_x^2 L(\bar{x},\bar{\lambda})]\Delta x + [h'(\bar{x})]^T \lambda^+ & = & -f'(\bar{x}) \\ h'(\bar{x})\Delta x & = & -h(\bar{x}) \end{array} \qquad (N)$$

$$\min_{\Delta x} \left\{ f(\bar{x}) + \Delta x^T f'(\bar{x}) \right. \qquad \text{(QP)}$$
$$\left. + \frac{1}{2}\Delta x^T \nabla_x^2 L(\bar{x},\bar{\lambda})\Delta x : h'(\bar{x})\Delta x = -h(\bar{x}) \right\}$$

**Proof** of Critical Observation: Let $z$ be a KKT points of (QP), and $\mu$ be the corresponding vector of Lagrange multipliers. The KKT system for (QP) reads

$$\begin{array}{rcl} f'(\bar{x}) + \nabla_x^2 L(\bar{x},\bar{\lambda})z + h'(\bar{x})\mu & = & 0 \\ h'(\bar{x})z & = & -h(\bar{x}) \end{array}$$

which are exactly the equations in $(N)$. Since the matrix of system $(N)$ is nonsingular, we have $z = \Delta x$ and $\mu = \lambda^+$.

$$\min_x \left\{ f(x) : h(x) \equiv (h_1(x), ..., h_k(x))^T = 0 \right\} \qquad (P)$$

♣ The Newton method as applied to the KKT system of $(P)$ works as follows:

*Given current iterate $(\bar{x}, \bar{\lambda})$, we linearize the constraints, thus getting "approximate feasible set"*

$$\bar{T} = \{\bar{x} + \Delta x : h'(\bar{x})\Delta x = -h(\bar{x})\},$$

*and minimize over this set the quadratic function*

$$f(\bar{x}) + (x-\bar{x})^T f'(\bar{x}) + \frac{1}{2}(x-\bar{x})^T \nabla_x^2 L(\bar{x}, \bar{\lambda})(x-\bar{x}).$$

*The solution of the resulting quadratic problem with linear equality constraints is the new $x$-iterate, and the vector of Lagrange multipliers associated with this solution is the new $\lambda$-iterate.*

Note: The quadratic part in the auxiliary quadratic objective comes from the Lagrange function of $(P)$, and not from the objective of $(P)$!

## General constrained case

♣ "Optimization-based" interpretation of the Newton method as applied to the KKT system of equality constrained problem can be extended onto the case of general constrained problem

$$\min_x \left\{ f(x) : \begin{array}{l} h(x) = (h_1(x), ..., h_k(x))^T = 0 \\ g(x) = (g_1(x), ..., g_m(x))^T \leq 0 \end{array} \right\} \qquad (P)$$

and results in the Basic SQP scheme:

*Given current approximations* $x_t$, $\lambda_t$, $\mu_t \geq 0$ *to a nondegenerate local solution* $x_*$ *of* $(P)$ *and corresponding optimal Lagrange multipliers* $\lambda^*$, $\mu^*$, *we solve auxiliary linearly constrained quadratic problem*

$$\min_{\Delta x} \left\{ f(x_t) + \Delta x^T f'(x_t) + \frac{1}{2}\Delta x^T \nabla_x^2 L(x_t; \lambda_t, \mu_t)\Delta x : \begin{array}{l} h'(x_t)\Delta t = -h(x_t) \\ g'(x_t)\Delta x \leq -g(x_t) \end{array} \right\}$$

$$L(x; \lambda, \mu) = f(x) + h^T(x)\lambda + g^T(x)\mu \qquad (QP_t)$$

*set* $x_{t+1} = x_t + \Delta x_*$ *and define* $\lambda_{t+1}$, $\mu_{t+1}$ *as the optimal Lagrange multipliers of* $(\mathrm{QP}_t)$*.*

$$\min_x \left\{ f(x) : \begin{array}{l} h(x) = (h_1(x), ..., h_k(x))^T = 0 \\ g(x) = (g_1(x), ..., g_m(x))^T \le 0 \end{array} \right\} \quad (P)$$

Theorem. Let $(x_*; \lambda^*, \mu^*)$ be a nondegenerate locally optimal solution to $(P)$ and the corresponding optimal Lagrange multipliers. The Basic SQP method, started close enough to $(x_*; \lambda^*, \mu^*)$, and restricted to work with appropriately small $\Delta x$, is well defined and converges to $(x_*; \lambda^*, \mu^*)$ quadratically.

♣ Difficulty: From the "global" viewpoint, the auxiliary quadratic problem to be solved may be bad (e.g., infeasible or below unbounded). In the equality constrained case, this never happens when we are close to the nondegenerate local solution; in the general case, bad things may happen even close to a nondegenerate local solution.

♣ Cure: replace the matrix $\nabla_x^2 L(x_t; \lambda^t, \mu^t)$ when it is not positive definite on the entire space by a positive definite matrix $B_t$, thus arriving at the method where the auxiliary quadratic problem is

$$\begin{array}{l}\min\limits_{\Delta x} \Big\{ f(x_t) + \Delta x^T f'(x_t) \\ +\frac{1}{2}\Delta x^T B_t \Delta x : \begin{array}{l} h'(x_t)\Delta t = -h(x_t) \\ g'(x_t)\Delta x \leq -g(x_t) \end{array} \Big\} \end{array} \tag{QP$_t$}$$

With this modification, the auxiliary problems are convex and solvable with unique optimal (provided that they are feasible, which indeed is the case when $x_t$ is close to a nondegenerate solution to $(P)$).

## Ensuring global convergence

♣ "Cured" Basic SQP scheme possesses nice local convergence properties; however, it in general is not globally converging.
Indeed, in the simplest unconstrained case SQP becomes the basic/modified Newton method, which is not necessarily globally converging, unless linesearch is incorporated.

♠ To ensure global convergence of SQP, we incorporate linesearch. In the scheme with linesearch, the solution $(\Delta x, \lambda^+, \mu^+)$ to the auxiliary quadratic problem

$$\min_{\Delta x} \left\{ f(x_t) + \Delta x^T f'(x_t) \right.$$
$$\left. + \frac{1}{2}\Delta x^T B_t \Delta x : \begin{array}{l} h'(x_t)\Delta t = -h(x_t) \\ g'(x_t)\Delta x \leq -g(x_t) \end{array} \right\} \tag{QP$_t$}$$

is used as *search direction* rather than as a new iterate. The new iterate is

$$\begin{array}{rcl} x_{t+1} & = & x_t + \gamma_{t+1}\Delta x \\ \lambda_{t+1} & = & \lambda_t + \gamma_{t+1}(\lambda^+ - \lambda_t) \\ \mu_{t+1} & = & \mu_t + \gamma_{t+1}(\mu^+ - \mu_t) \end{array}$$

where $\gamma_{t+1} > 0$ is the stepsize given by linesearch.

Question: What should be minimized by the linesearch?

♣ In the constrained case, the auxiliary objective to be minimized by the linesearch cannot be chosen as the objective of the problem of interest. In the case of SQP, a good auxiliary objective ("merit function") is

$$M(x) = f(x) + \theta\left[\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{k} g_j^+(x)\right]$$

where $\theta > 0$ is parameter.

Fact: Let $x_t$ be current iterate, $B_t$ be a positive definite matrix used in the auxiliary quadratic problem, $\Delta x$ be a solution to this problem and $\lambda \equiv \lambda_{t+1}$, $\mu \equiv \mu_{t+1}$ be the corresponding Lagrange multipliers. Assume that $\theta$ is large enough:

$$\theta \geq \max\{|\lambda_1|, ..., |\lambda_k|, \mu_1, \mu_2, ..., \mu_m\}$$

Then either $\Delta x = 0$, and then $x_t$ is a KKT point of the original problem, or $\Delta x \neq 0$, and then $\Delta x$, that is,

$$M(x + \gamma\Delta x) < M(x)$$

for all small enough $\gamma > 0$.

## SQP Algorithm with Merit Function

♣ Generic SQP algorithm with merit function is as follows:

♢Initialization: Choose $\theta_1 > 0$ and starting point $x_1$

♢Step $t$: Given current iterate $x_t$,

— choose a matrix $B_t \succ 0$ and form and solve auxiliary problem

$$\min_{\Delta x} \Big\{ f(x_t) + \Delta x^T f'(x_t) + \frac{1}{2}\Delta x^T B_t \Delta x : \begin{array}{l} h'(x_t)\Delta t = -h(x_t) \\ g'(x_t)\Delta x \leq -g(x_t) \end{array} \Big\} \qquad (\text{QP}_t)$$

thus getting the optimal $\Delta x$ along with associated Lagrange multipliers $\lambda, \mu$.

— if $\Delta x = 0$, terminate: $x_t$ is a KKT point of the original problem, otherwise proceed as follows:

— check whether

$$\theta_t \geq \bar{\theta}_t \equiv \max\{|\lambda_1|, ..., |\lambda_k|, \mu_1, ..., \mu_m\}.$$

if it is the case, set $\theta_{t+1} = \theta_t$, otherwise set

$$\theta_{t+1} = \max[\bar{\theta}_t, 2\theta_t];$$

— Find the new iterate

$$x_{t+1} = x_t + \gamma_{t+1}\Delta x$$

by linesearch aimed to minimize the merit function

$$M_{t+1}(x) = f(x) + \theta_{t+1}\left[\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{k} g_j^+(x)\right]$$

on the search ray $\{x_t + \gamma\Delta x \mid \gamma \geq 0\}$. Replace $t$ with $t+1$ and loop.

$$\min_x \left\{ f(x) : \begin{array}{l} h(x) = (h_1(x), ..., h_k(x))^T = 0 \\ g(x) = (g_1(x), ..., g_m(x))^T \leq 0 \end{array} \right\} \quad (P)$$

<u>Theorem:</u> Let general constrained problem be solved by SQP algorithm with merit function. Assume that

• there exists a compact $\Omega \subset \mathbf{R}^n$ such that for $x \in \Omega$ the solution set $D(x)$ of the system of linear inequality constraints

$$S(x) : \quad h'(x)\Delta x = -h(x),\ g'(x)\Delta x \leq -g(x)$$

with unknowns $\Delta x$ is nonempty, and each vector $\Delta x \in D(x)$ is a regular solution of system $S(x)$;

• the trajectory $\{x_t\}$ of the algorithm belongs to $\Omega$ and is infinite (i.e., the method does not terminate with exact KKT point);

• the matrices $B_t$ used in the method are uniformly bounded and uniformly positive definite: $cI \preceq B_t \preceq CI$ for all $t$, with some $0 < c \leq C < \infty$.

Then all accumulation points of the trajectory of the method are KKT points of $(P)$.