Large-Scale Optimization and Nonlinear Systems of Equations

Typical problem and some classification

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{cases} c_i(x) = 0 & i \ \mathcal{E} \\ c_i(x) \geq 0 & i \ \mathcal{I} \end{cases}$$

x vector of variables, parameters, unknowns

f objective function (scalar)

c_i constraint functions

$\mathcal{E}$ set of equality constraints

$\mathcal{I}$ set of inequality constraints

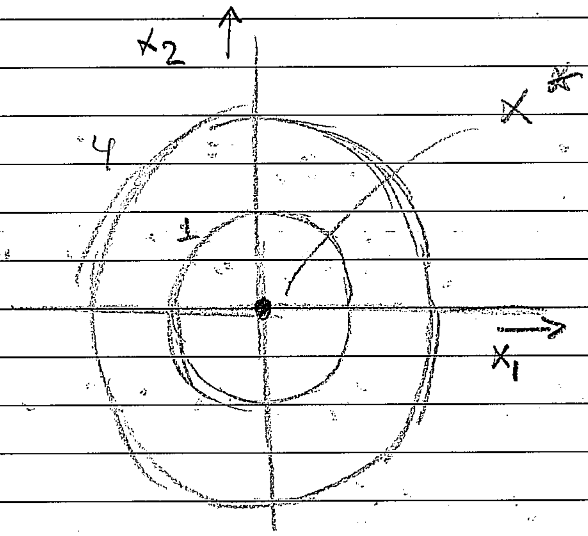
Set of points satisfying all constraints:

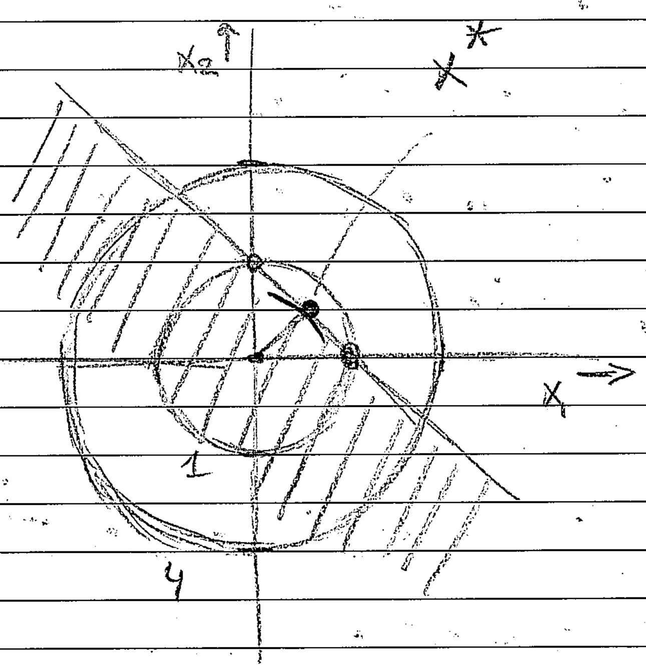
feasible region / feasible set

x^* solution

In principle, we consider continuous optimization →

all components x_i are real (or complex).


x_2
4
1
x^*
x_1


x_2
x^*
x_1
1
4

→ If all variables x_i are discrete (say integers) we have a discrete optimization problem. If some are discrete and some continuous, we call the problem a mixed (integer) programming problem. In discrete opt. x is an element of a finite (but often very large) set.

→ This course focuses on continuous optimization probs

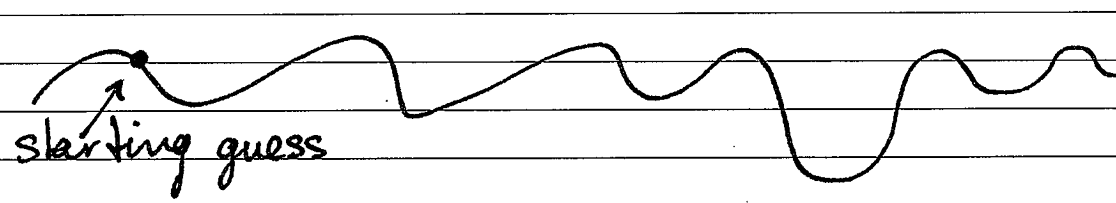
→ Opt. problems can be constrained or unconstrained. We consider both.

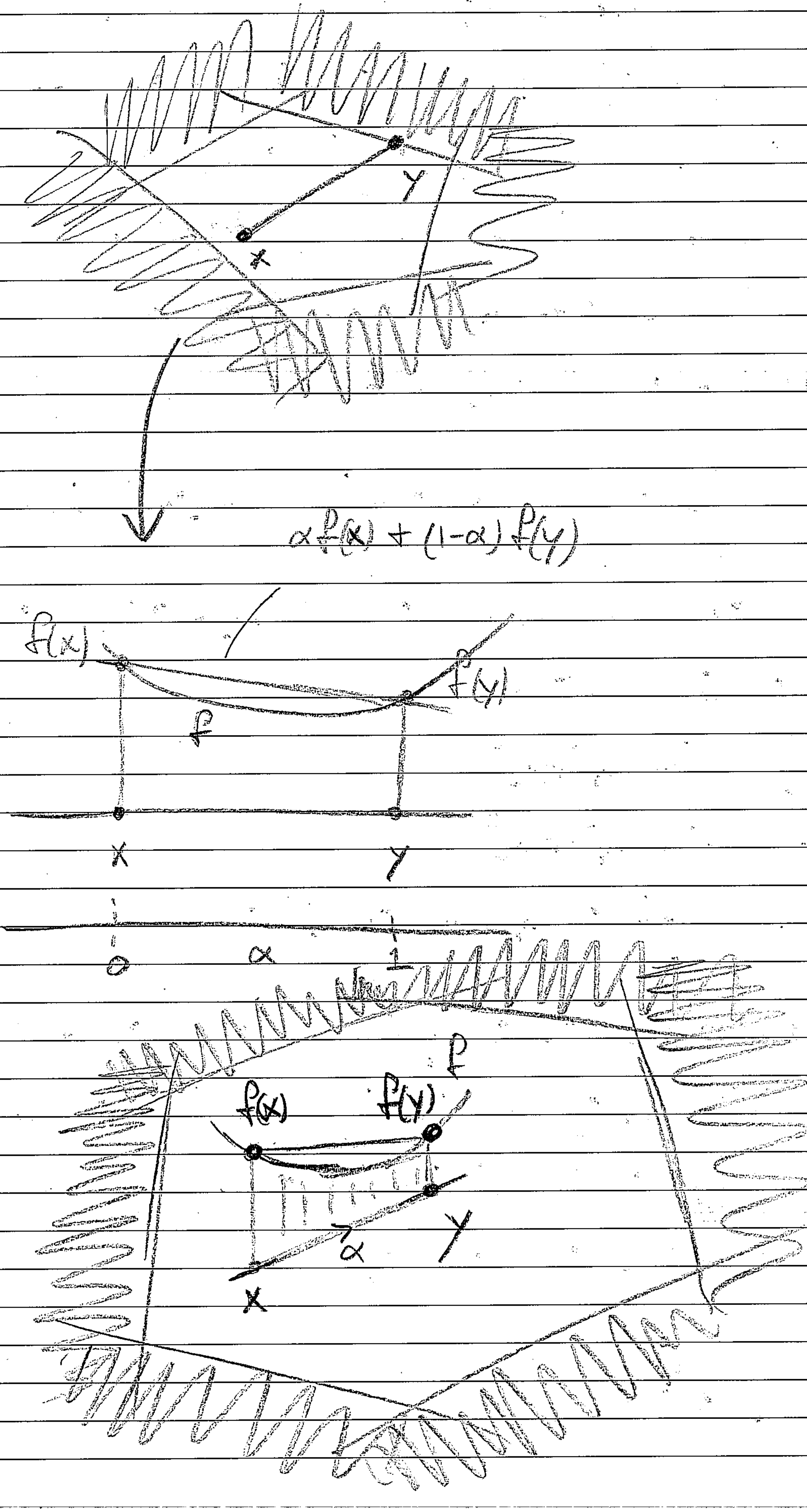
a) $\min \; x_1^2 + x_2^2$

b) $\min \; x_1^2 + x_2^2$

subj. to $x_1 + x_2 - 1 \geq 0$

→ Although we often/in general want global minimum, often finding this is too expensive (many local minima, no clear path to get from one to another)




y
x
αf(x) + (1-α) f(y)
f(x)
f(y)
f
x
y
0
α
1
f
f(x)
f(y)
y
α
x

In most cases, we settle for <u>local</u> <u>min</u>.
(obj. func. smaller than for all other ~~poin~~ feasible points in neighborhood)

For <u>convex</u> <u>progr.</u> <u>problems</u> local solutions are global solutions. This is particularly the case for <u>linear</u> <u>programming</u> problems

Convex problems are easier than non-convex problems.

$S \subseteq \mathbb{R}^n$ <u>convex</u> <u>set</u> if

$\forall x,y \in S : \alpha x + (1-\alpha) y \in S$ for all $\alpha \in [0,1]$

(any point on straight line segment between $x \in S$ and $y \in S$ is element of S)

$f: S \subseteq \mathbb{R}^n \to \mathbb{R}$ is convex function if domain S convex set and

$\forall x,y \in S : f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$
for all $\alpha \in [0,1]$

f strictly convex if inequality is strict whenever $x \neq y$ and $\alpha \in (0,1)$

f <u>concave</u> if −f convex.

→ If objective function and feasible region are both convex, then a local solution is also the global solution.

The term convex programming describes a special case of the general constrained optimization problem where:

i) the objective function is convex

ii) the equality constraints are linear

iii) the inequality constraint functions are concave.

→ Problems can be <u>stochastic</u> or <u>deterministic</u>. We focus on determ. problems, in which the model is completely known.

→ For practical algorithms, we focus on

- x robustness
- x efficiency
- x accuracy

Basics (charac. solutions / algorithms)

$\min_x f(x) \quad x \in \mathbb{R}^n \quad (n \geq 1)$, $f: \mathbb{R}^n \to \mathbb{R}$ (smooth)

f expensive, fewer function evaluations efficient

Example

$$\phi(t; x) = x_1 + x_2 e^{-(x_3 - t)^2 / x_4} + x_5 \cos(x_6 t)$$

x_i parameters, y_j data at t_j $(j = 1..m)$

residual $r_j(x) = y_j - \phi(t_j; x)$

$$\min_{x \in \mathbb{R}^6} f(x) = \sum_{j=1}^{m} r_j^2(x) \qquad \left(\text{or } \tfrac{1}{2}\Sigma \cdots\right)$$

→ nonlinear least squares problem

other norms (than $\| \cdots \|_2$) possible too

$$\min_{x \in \mathbb{R}^6} \| r(x) \|$$

possibly constrained: $\min_{\substack{x \in \mathbb{R}^6 \\ x_i \geq 0}} \| r(x) \|$

Best global minimizer of f

x^* is global minimizer if $f(x^*) \leq f(x)$ for all x $(x \in \mathbb{R}^n$ or $x \in D \subset \mathbb{R}^n)$

Global minimizer hard to find (very expensive), so usually local minimizer

x^* is local minimizer if there is a neighborhood N of x^* s.t. $f(x^*) \le f(x)$ for all $x \in N$ (also weak local minimizer, strict/strong local minimizer if $f(x^*) < f(x)$ for all $x \in N$ and $x \neq x^*$)

Example: constant function every point (weak) local min

x^* isolated local min if n.hood N exists s.t. x only local min in N

Strict local min need not be isolated *
Isolated local min always strict

* inf seq. of min x_j and $x_j \to \hat{x}$ then $\hat{x}$ not isolated (even if strict)

$x_j = 2^{-j}$

Analyze (candidate) solutions based on derivatives (if f suff. smooth)

Theo 2.1 / Taylor's Theo

Let $f: \mathbb{R}^n \to \mathbb{R}$ be cont. diff. and $p \in \mathbb{R}^n$

Then

$$f(x+p) = f(x) + \nabla f(x+tp)^T p$$

for some $t \in (0,1)$

If f twice cont. diff.,

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\, p \, dt$$

and

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp)\, p$$

for some $t \in (0,1)$

Necessary cond. for optimality :

First order necessary cond. , Theo 2.2

If x^* is local min and f cont. diff. in open Neighborhood of x^*, then $\nabla f(x^*) = 0$

Proof Assume $\nabla f(x^*) \neq 0$. Define $p = -\nabla f(x^*)$.

Then $p^T \nabla f(x^*) = -p^T p = -\|\nabla f(x^*)\|_2^2 < 0$

Because ∇f cont. in n.hood of x^*, there is $T > 0$ s.t. $p^T \nabla f(x^* + tp) < 0$ forall $t \in [0, T]$.

Then, for all $\bar{t} \in (0, T]$, we have (Taylor's Theo)

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x + t\bar{t}p) \text{ for some } t \in (0,1)$$

Hence $f(x^* + \bar{t}p) < f(x^*)$ for all $t \in (0, T]$

So, x^* is not local minimizer, and assump. leads to a contradiction (hence must be false)

x^* is stationary point if $\nabla f(x^*) = 0$

Theo 2.2 any local min must be stat. pt.

(here typ. symm.)

Matrix $B \in \mathbb{R}^{n \times n}$ positive definite if

$p^T B p > 0$ for all $p \neq 0$ and positive semidefinite if $p^T B p \geq 0$ for all p

Theo 2.3 (second order necessary cond.s)

If x^* is local min of f and $\nabla^2 f$ exists and is continuous in open n.hood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Proof: Theo 2.2 $\Rightarrow \nabla f(x^*) = 0$. Assume that $\nabla^2 f(x^*)$ not pos. semidefinite. Then p exists s.t. $p^T \nabla^2 f(x^*) p < 0$ and because $\nabla^2 f$ cont near x^*, there is scalar $T > 0$ s.t.

$p^T \nabla^2 f(x^* + tp) p < 0$ for all $t \in [0, T]$.

Using a Taylor series expansion around x^*

we have for all $\bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$ that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^*) + \frac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp)p < f(x^*) .$$

~~[illegible]~~

So, x^* is not local minimizer

<u>Sufficient</u> conditions for local minimizer

<u>Theo 2.4</u> (second order suff. cond.s)

Suppose that $\nabla^2 f$ is cont. in an open n.hood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x)$ is positive definite. Then x^* is strict local min of f.

<u>Proof</u>: Because Hessian ($\nabla^2 f(x)$) is continuous and positive definite at x^*, we can choose radius $r > 0$ so that $\nabla^2 f(x)$ is positive definite for all x in the open ball

$$D = \{ z : \|z - x^*\|_2 < r \}$$

So, for any p with $\|p\| < r$, we have $x^* + p \in D$ and so

$$f(x^* + p) = f(x^*) + \underbrace{p^T \nabla f(x^*)}_{=0} + \frac{1}{2} p^T \nabla^2 f(z) p =$$

$$f(x^*) + \frac{1}{2} p^T \nabla^2 f(z) p$$

where $z = x^* + tp$ for some $t \in (0,1)$.

Since $z \in D$, $p^T \nabla^2 f(z) p > 0$ and therefore $f(x^*+p) > f(x^*)$. (which proves desired result)

(discuss)

Note that there is a "gap" between necessary conditions and sufficient cond.s (here)

There are solutions (even strict local min) that do not satisfy Theo 2.4.

There are cases satisfying Theo 2.2/2.3 that do not give a local min.

(semidef. Hessian with (say) with higher directional deriv. negative)

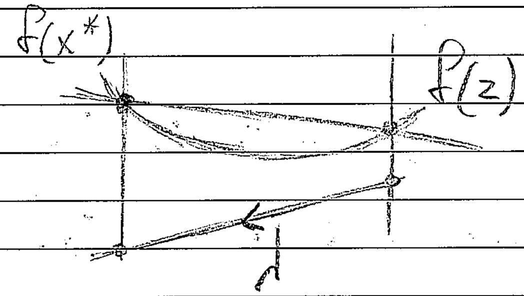
Theo 2.5

When f is convex, any local minimizer x^* is a global minimizer of f. If in addition f is differentiable, then any stationary point x^* is a global minimizer of f.
$(x \in \mathbb{R}^n)$

Proof: Suppose x^* is local but not global min. Then there exists a point $z \in \mathbb{R}^n$ s.t. $f(z) < f(x^*)$. Consider line segment that joins x^* to z: $x = \lambda z + (1-\lambda)x^*$ (for some $\lambda \in (0,1]$)
By convexity of f:

$$f(x) \leq \lambda f(z) + (1-\lambda) f(x^*) < f(x^*)$$

(because $f(z) < f(x^*)$
$\lambda f(z) + (1-\lambda) f(x^*)$ strictly increasing)



Any neighborhood N of x^* contains part of the line segment $x = \lambda z + (1-\lambda)x^*$ for $\lambda \in (0,1]$. So, N will contain points x where $f(x) < f(x^*)$

Hence x^* is not a local min. (so assump. leads to a contradiction)

2nd part

~~Suppose ∇x^* is a stat. point~~

Let x^* be a stationary point. ($\nabla f(x^*) = 0$).
Assume x^* is not global min. Then there is a z s.t. $f(z) < f(x^*)$. Consider same line segment as above, and the deriv. of f along this segment (wrt λ) at x^*

$$\nabla f(x^*)^T (z - x^*) = \frac{d}{d\lambda} f(x^* + \lambda(z - x^*)) \Big|_{\lambda=0} =$$

$$\lim_{\lambda \downarrow 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \leq$$

$$\lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1-\lambda) f(x^*) - f(x^*)}{\lambda} =$$

$$f(z) - f(x^*) < 0$$

Therefore $\nabla f(x^*) \neq 0$ (contradiction)

~~So x^* is not~~ So, x^* must be global min.

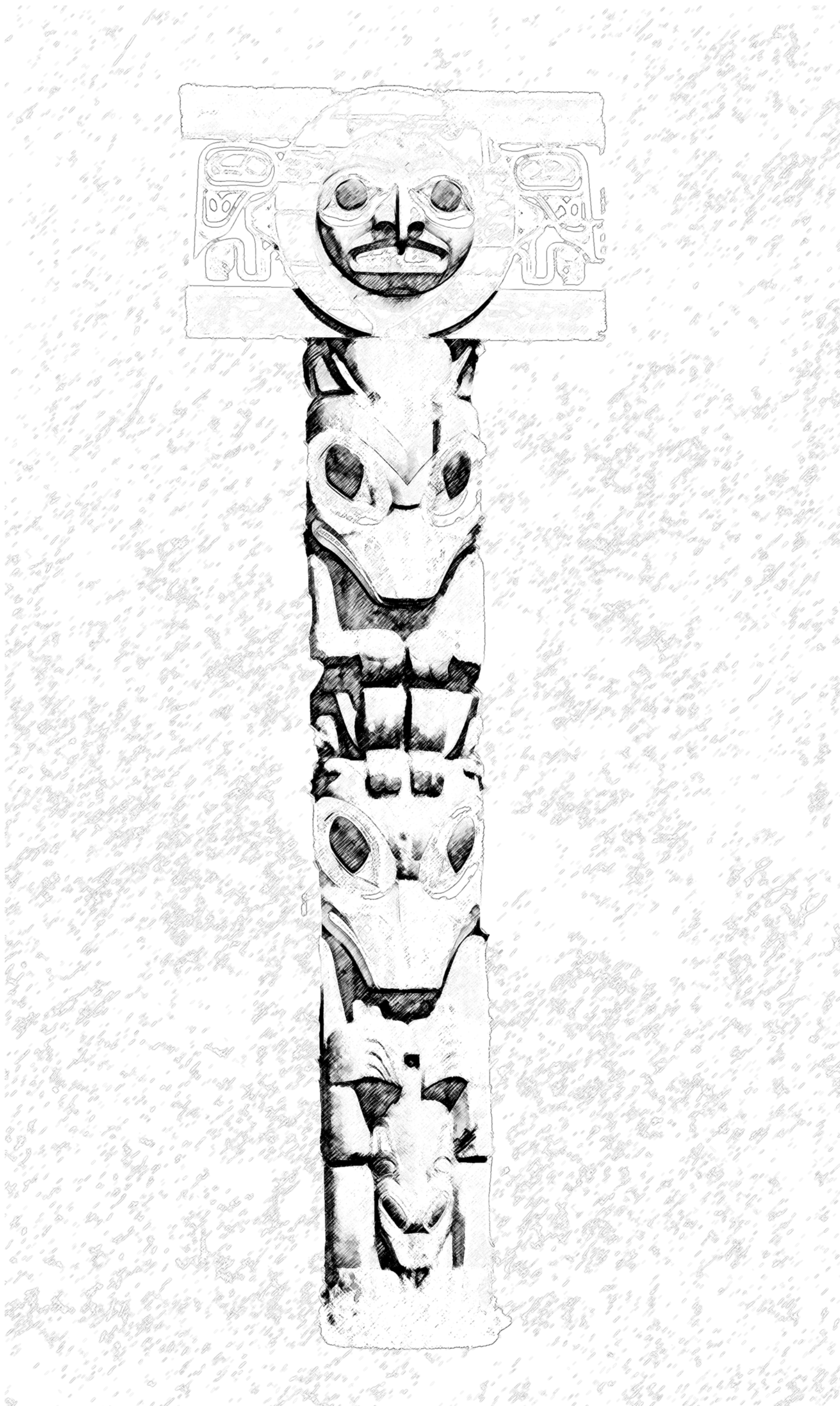
(book: Let x^* not be glob. min.

Assume f stationary at x^* ...

contradiction $\rightarrow$ f not stat. at x^*

13

Most alg.s are based on finding stat. points
(and then checking add. cond.s)



Overview of Algorithms

Many alg.s, describe most common general form (unconstrained)

Initial guess x_0

- good guess by knowl. user or based on previous computations
- arbitrary guess (if no prior info)
- some procedure to get started
  (e.g. some multilevel or sampling approach)

Generate seq. iterates $\{x_k\}_{k=0}^{\infty}$

Stop when suff. acc. solution or lack of progress (robustness)

Use info from previous iterate (or modest nr. of iterates) and local info from that/those iterate(s) to compute next point

Since most appr. are based on local info (from few points at most) robustness requires that we don't jump to far.

Two strategies: Line Search & Trust Region

<u>Line search</u>: given position x_k, choose/find direction p_k and find good stepsize in that direction (α):

$$\min_{\alpha > 0} f(x_k + \alpha p_k)$$

In general no need to solve min-zation exactly. In fact, often undesirable (other considerations) ↳ later

L.S. alg. generates modest nr. of trial steps until it finds acceptable one: $x_{k+1} = x_k + \alpha p_k$

<u>Trust region</u>: define (local) model function, m_k, based on local info about f (deriv.s):

$$\min_{p \in \mathbb{R}^n} m_k(x_k + p) \quad \text{where } x_k + p \text{ inside "trust region"}$$

→ region where we trust local model
→ region updated every step depending on how well result from opt. model m_k corresponds to f_k

→ example $x_k + p \in B(x_k, r)$
↳ $\|p\| \leq r$

Insuff. decrease in f: shrink trust region and resolve (local info → better updates, relative to size, for smaller region)

If predicted improvement from m_k very close to actual improvement (from $f(\cdot)$) $\rightarrow$ increase trust region (allow bigger steps)

Typically $m_k(x_k+p) = f_k + p^T\nabla f_k + \frac{1}{2}p^T B_k p$

where $f_k = f(x_k)$, $\nabla f_k = \nabla f(x_k)$ and B_k is an approximate Hessian of f at x_k (or exact Hessian $\nabla^2 f(x_k)$)

Workout example p. 19/20 (graphs in Matlab)

Trust region : both direction and step length change, in general, if size trust region is changed

Line search: direction chosen, then fixed. Only step length varies.

<u>Search directions for Line Search</u>

Need to find directions to decrease f_k $(= f(x_k))$

Obvious choice $-\nabla f(x_k)$ → direction of fastest decrease : $p_k = -\nabla f(x_k) / \|\nabla f(x_k)\|$

↳ $\|p_k\| = 1$

Find good α for $f(x_k + \alpha p_k)$

If $f(\cdot)$ expensive, trying many α's too expens.

→ (Method of) Steepest Descent

In general, any <u>descent direction</u> is possible (any direction s.t. $p^T \nabla f(x_k) < 0$ <u>or</u> $\angle(p, -\nabla f(x_k)) < \pi/2$)

Any <u>descent direction</u> gives improvement for small enough step.

$$f(x_k + \varepsilon p_k) = f(x_k) + \varepsilon p_k^T \nabla f_k + \frac{1}{2}\varepsilon^2 p_k^T \nabla f(x_k + t\varepsilon p_k) p_k$$

for some $t \in (0,1)$

$$= f_k + \varepsilon p_k^T \nabla f_k + O(\varepsilon^2)$$ [illegible]

$p_k^T \nabla f(x_k + t\varepsilon p_k) p$ bounded [illegible]
(f twice cont. diff.)

$$p_k^T \nabla f_k = \|p_k\| \|\nabla f_k\| \cos\theta_k < 0 \quad (\|p_k\| = 1)$$

where $\theta_k < 0$

So, for ε suff. small $f(x_k + \varepsilon p_k) < f(x_k)$

[illegible]

Alternative Newton direction

$$f(x_k + p) \simeq f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p \overset{\text{def}}{=} m_k(p)$$

Minimize model $m_k(p)$ (assume $\nabla^2 f_k$ pos. def.)

$$\rightarrow p^N = -(\nabla^2 f_k)^{-1} \nabla f_k$$

$m(p) = \frac{1}{2} p^T A p + p^T b + c$ $\quad$ A SPD

consider p minimizer. Then for any direction q the derivative in direction q must vanish (otherwise we can reduce m)

$$\frac{d}{d\varepsilon}(m + \varepsilon q)\Big|_{\varepsilon=0} = 0 \quad \underline{\text{any}}\ q \neq 0 \ (\text{for all } q \neq 0)$$

$$\lim_{\varepsilon \downarrow 0} \frac{d}{d\varepsilon} \frac{1}{2}(p+\varepsilon q)^T A (p + \varepsilon q) + (p + \varepsilon q)^T b + c$$

$$= \lim_{\varepsilon \downarrow 0} \frac{d}{d\varepsilon}\left(\frac{1}{2} p^T A p + \frac{1}{2}\varepsilon^2 q^T A q + \varepsilon q^T A p + p^T b + \varepsilon q^T b + c\right)$$

$$= \lim_{\varepsilon \downarrow 0} \varepsilon q^T A q + q^T A p + q^T b = 0$$

$$q^T (A p + b) = 0 \quad \text{for all } q \ (q \neq 0)$$

$$\Rightarrow A p + b = 0 \Rightarrow p = A^{-1} b$$

A^{-1} exists if A SPD

Newton direction is reliable (good search direction) if $f(x_k + p) \simeq m_k(p)$ (more precise later)

We replaced $\nabla^2 f(x_k + tp)$ by $\nabla^2 f(x_k)$ which is okay if p not too large and $\nabla^2 f$ does not change too quickly

If $\nabla^2 f_k$ pos. def. then Newton direction is search direction

$$(p_k^N)^T \nabla f_k = - \nabla f_k^T (\nabla^2 f_k)^{-1} \nabla f_k < 0$$

by def. $z^T A z > 0$ $(z \neq 0)$ if A positive def.

what about $z^T A^{-1} z$?

let $z = Av$ (A cannot be singular)

$$z^T A^{-1} z = v^T A^T A^{-1} A v = v^T A^T v = v^T A v > 0$$

(real case)

$$\nabla f_k^T p_N = -(p_N)^T \nabla^2 f_k^T p_N = -p_N^T \nabla^2 f_k p_N$$

$$\leq -\sigma \|p_N\|^2$$

for some positive σ

So, assuming $\nabla f_k \neq 0$ (we would be done)

p_N is descent direction

In principle no reason to take other step than whole Newton step p_N ($\rightarrow \alpha = 1$)

However, based on local model, so possibly bad if (in some relative sense) $\|p\|$ to large.

In that case reduce α → <u>damping</u> or <u>backtracking</u>.

If $\nabla^2 f_k$ not positive definite problem

a) $\nabla^2 f_k$ singular : no Newton step
(if nearly singular → step very large)

b) $-(\nabla^2 f_k)^{-1} \nabla f_k = p_N$ not a descent direction

Typically, modify search ~~step~~ direction

Advantage of Newton is fast <u>local conv.</u> (close enough to sol.), typically quadratic.
Disadvantage cost of comp. $\nabla^2 f(x_k)$, often expensive (sometimes imposs. to comp. exactly)
→ approx. by finite diff. or direct. diff.

Quasi Newton methods: replace Hessian $\nabla^2 f_k$ by approx. B_k (typically updated)

For example: use change in gradient for info on second derivatives

Taylor: $\nabla f(x+p) = \nabla f(x) + \nabla^2 f(x) p + \underbrace{\int_0^1 [\nabla^2 f(x+tp) - \nabla^2 f(x)] p \, dt}_{o(\|p\|)}$

$p = x_{k+1} - x_k \qquad (x_{k+1} = x_k + p)$

$\nabla f_{k+1} = \nabla f_k + \nabla^2 f_k (x_{k+1} - x_k) + \underbrace{o(\|p\|)}_{o(\|x_{k+1} - x_k\|)}$

So, $\nabla^2 f_k (x_{k+1} - x_k) \simeq \nabla f_{k+1} - \nabla f_k$

If Iteration $x_k, x_{k+1}, \dots$ converges to solution x^* then this approx. gets more and more accurate (assuming $\nabla^2 f$ pos. def. at the solution)

Update B_{k+1} (from B_k) to mimic property

Secant condition: $B_{k+1} s_k = y_k$ where

$s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$

In optimization also require B_{k+1} symm. (like Hessian)

Update cheap → low rank

Symmetric-rank-one (SR1) update:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

BFGS formula/update (Broyden, Fletcher, Goldfarb, Shanno)

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$ (rank-two update)

(note B_k already symmetric)

$$p_k = - B_k^{-1} \nabla f_k$$

BFGS maintains pos. def. if B_k pos def and $s_k^T y_k > 0$

To avoid ~~sa~~ solving system & work with $H_k = B_k^{-1}$

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k s_k y_k^T) + \rho_k s_k s_k^T$$

where $\rho_k = (y_k^T s_k)^{-1}$

$$p_k = - H_k \nabla f_k$$

Variants ~~discuss~~ discussed later.

Alternative quasi-Newton:

nonlinear conjugate gradient methods

$p_k = -\nabla f(x_k) + \beta_k p_{k-1}$

β_k scalar enforcing ~~&~~ p_k, p_{k-1} conjugate
~~(orthog. with inner prod. defined by Hessian)~~ / ~~orthog~~

orthog. condition (forcing search directions to "stay away" from each other)

Near solution corr. to minimizing

$\phi(x) = \frac{1}{2} x^T A x - x^T b$

A Hessian at sol.

$\nabla \phi(x) = Ax - b \quad (= -r \text{ of } Ax = b)$

Models for Trust-Region Methods

If $B_k = 0$ (choice) in (and region by Eucl. norm)

$$m_k(x_k+p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p \quad \rightarrow$$

$$\min_p \; f_k + p^T \nabla f_k \quad \text{subject to} \quad \|p\|_2 \le \Delta_k$$

$\rightarrow$

$$p = - \frac{\nabla f_k}{\|\nabla f_k\|} \cdot \Delta_k$$

Steepest descent step with step length gov'd by trust region radius. (similar to line search)

Another common choice (Newton method) is

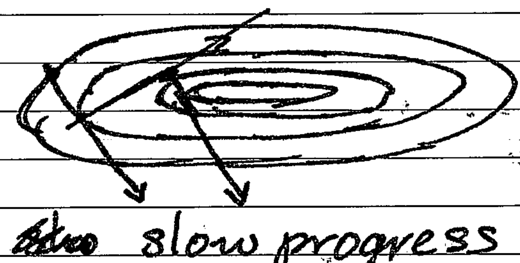
$$B_k = \nabla^2 f_k$$

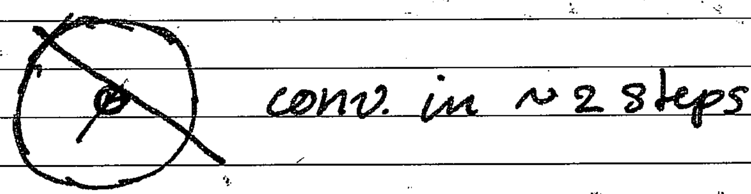
This choice gives solution even if Hessian not positive definite because of trust region constraint.

If B_k is def.d by a quasi-Newton approximation we obtain a trust-region quasi-Newton method.

Poor scaling of problem means f much more sensitive (higher rate of change) in some direction than in another.

Poor scaling tends to hinder fast convergence → try to rescale problem or carry out steps wrt judiciously chosen norms.





~~Most alg.s are based on finding stat points (and then checking add. cond.s)~~

Basic Analysis

→ after intro basics

~~Where~~ (from appendix A2)

let $\{x_k\}$ be (infinite) seq. of points in $\mathbb{R}^n$

$\{x_k\}$ converges to x : $\lim_{k \to \infty} x_k = x$ if for any $\varepsilon > 0$, there is an index K s.t.

$\|x_k - x\| \leq \varepsilon$ for all $k \geq K$

Given an (ordered) index set $S \subset \{1, 2, 3, \dots\}$ we can def. a subsequence of $\{x_k\}$ corr. to S ($\{x_k\}_{k \in S}$).

We say $\hat{x}$ is an accumulation point or limit point for $\{x_k\}$ if there is an infinite set of indices $k_1, k_2, \dots$ s.t. subsequence $\{x_{k_i}\}_{i=1,2,\dots}$ converges to $\hat{x}$: $\lim_{i \to \infty} x_{k_i} = \hat{x}$

(alternative: for any $\varepsilon > 0$ and all positive integers K, we have $\|x_k - \hat{x}\| \leq \varepsilon$ for some $k \geq K$ (at least one))

* A seq. can have infinite number of limit pts
* A seq. converges if it has exactly one limit pt.

A seq. is Cauchy seq if for any $\varepsilon > 0$, there exists integer K s.t. $\|x_k - x_\ell\| < \ell$ for all $k, \ell \geq K$

In $\mathbb{R}^n$ a seq. conv. if and only if it is a Cauchy seq.

For a scalar seq. $\{t_k\}$ ($t_k \in \mathbb{R}$), seq is bounded above if there exists $u \in \mathbb{R}$ s.t. $t_k \leq u$ for all k

bounded below if there exists $v \in \mathbb{R}$ s.t. $t_k \geq v$ for all k

$\{t_k\}$ non-decreasing if $t_{k+1} \geq t_k$ for all k

$\{t_k\}$ non-increasing if $t_{k+1} \leq t_k$ $\forall k$

<u>Theo</u>: If $\{t_k\}$ is non-decreasing and bounded above, then it converges

($\lim_{k\to\infty} t_k = t$ for some $t \in \mathbb{R}$)

(likewise for non-increasing and bounded below, $\{t_k\}$ converges)

Supremum of $\{t_k\}$ is smallest real number $u \in \mathbb{R}$ s.t. $t_k \leq u$ for all k. Write $\sup\{t_k\}$

Infimum of $\{t_k\}$ is largest real number v s.t. $v \leq t_k$ for all k $\rightarrow \inf\{t_k\}$

Define $u_i = \sup\{t_k \mid k \geq i\}$

Clearly $\{u_i\}$ is non-increasing. If bounded below, it converges to $\bar{u} \in \mathbb{R}$, called $\limsup t_k$

Define $v_i = \inf\{t_k \mid k \geq i\}$. Non-decreasing. If bnd above, it converges to $\bar{v} \in \mathbb{R}$ called $\liminf t_k$

$\{t_k\} = \{1, \frac{1}{2}, 1, \frac{1}{4}, 1, \frac{1}{8}, \cdots\}$

$\liminf t_k = 0$

$\limsup t_k = 1$

Rates of Convergence

$\{x_k\}$ seq. in $\mathbb{R}^n$ $\quad x_k \to x^*$

conv. Q - linear if there is constant $r \in (0,1)$ s.t. $\dfrac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \le r$ for all k suff. large

distance to x^* improves each step by at least factor r bounded away from 1

Q → quotient

conv. Q superlinear if $\displaystyle\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$

ex. $1 + k^{-k}$

Q - quadratic convergence if there is a positive constant M (poss. M>1) s.t.

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \le M \quad \text{for } k \text{ suff. large}$$

similarly $\dfrac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \le M$

$(p > 1)$ → Q-order of convergence p

Weaker form of convergence

R-linear (R for root) if there is a seq. of non neg scalars ($\in \mathbb{R}$) $\{v_k\}$ s.t.

$\|x_k - x^*\| \le v_k$ for all k, and

$\{v_k\}$ converges Q-linearly to 0

seq. $\{\|x_k - x^*\|\}$ is dominated by $\{v_k\}$

Ex : $x_k = \begin{cases} 1 + (1/2)^k & k \text{ even} \\ 1 & k \text{ odd} \end{cases}$

→ $x_0 = 2$, $x_1 = 1$, $x_2 = 1\frac{1}{4}$, $x_3 = 1$, $x_4 = 1\frac{1}{16}$, ...

Note seq. not non-increasing (goes up every other step. Still

$|x_k - x^*| \le (1/2)^k \qquad (x^* = 1)$

and $(1/2)^k$ conv. Q-linearly to zero,

and $(1/2)^k$ dominates $\{|x_k - x^*|\}$.

So, $\{|x_k - x^*|\}$ conv. R-linearly to zero

or $\{x_k\}$ conv. R-lin to $1 \ (= x^*)$

$\{x_k\}$ conv. R-superlinearly to x^* if $\{\|x_k - x^*\|\}$ is dominated by $\{v_k\}$ converging Q-superlinearly to zero.

Analogous for R-quadratic conv.

Topology of Euclidean Space $\mathbb{R}^n$

Set F ($\subset \mathbb{R}^n$) is bounded if $M \in \mathbb{R}$ ($M > 0$) exists s.t. $\|x\| \le M$ for all $x \in F$

Subset $F \subset \mathbb{R}^n$ is open if for every $x \in F$ there exists $\varepsilon > 0$ s.t. $\overline{B(x,\varepsilon)} = \{y \in F \mid \|x - y\| \le \varepsilon\}$ is contained in F : $\overline{B(x,\varepsilon)} \subset F$

Set F is closed if for all possible seq.s of points $\{x_k\}$ in F, all limit points of $\{x_k\}$ are elements of F.

(alt: a set is closed if its complement in $\mathbb{R}^n$ is open)

Ex. $F = (0,1) \cup (2,3)$ is open

$F = [0,1] \cup [2,3]$ is closed

$F = (0,1]$ is neither open nor closed

$F = \mathbb{R}$ is open <u>and</u> closed

The interior of F, int F, is the largest <u>open</u> set contained in F

The closure of F is the smallest <u>closed</u> set containing F : cl F or $\overline{F}$

(alt. closure of F is union of F with set of all limit points of F)

$x \in \text{cl } F$ if $\lim_{k \to \infty} x_k = x$ for some seq.

$\{x_k\}$ of points in F $(x_k \in F)$

$F = (-1,1] \cup [2,4)$ then

$\text{cl } F = [-1,1] \cup [2,4]$, $\text{int } F = (-1,1) \cup (2,4)$

The union of finitely many closed sets is closed. The intersection of arbitrarily many closed sets is closed.

The intersection of finitely many open sets is open. The ~~a~~ union of arb. many open sets is open.

The set F is compact if every seq. $\{x_k\}$ (where $x_k \in F$ for all k) in F has at least one limit point, and all limit points are in F.

(every seq. in F contains a conv. subseq. in F)

$F \subset \mathbb{R}^n$ is closed and bounded $\Rightarrow$

F is compact

Given $x \in \mathbb{R}^n$, we call $N \subset \mathbb{R}^n$ a ~~[illegible]~~ neighborhood of x if it is an open set containing x

Examples

Union of infin. many closed sets may not be a closed set:

$L_k = [0, 1-(1/2)^k]$ (L_k closed)

$\bigcup_{k=1}^{\infty} L_k \equiv L$ not closed

Consider seq $x_k = 1 - \frac{1}{2}^k$. $x_k \in L_k \Rightarrow x_k \in L$

However $x_k \to 1$ (limit pt) but

$\forall k : 1 \notin L_k \Rightarrow 1 \notin L$.

So, there exists a seq. in L with limit pt not in L

Intersection of inf. many open sets may not be open set.

$L_k = (0, 1+(\frac{1}{2})^k)$ (L_k open)

$U_k L_k \equiv L$ not open

$\forall k : 1 \in L_k \Rightarrow 1 \in L$

However, any $B(1,\varepsilon)$ contains points that are not in L_k for suff. large k.

In particular $B(x,\varepsilon) = \{ y \in \mathbb{R}^n \mid \|x-y\| < \varepsilon \}$
→ open ball of radius ε around x.
is a n.hood of x $(\varepsilon > 0)$

Given $F \subset \mathbb{R}^n$, we say N is n.hood of F if there is $\varepsilon > 0$ s.t. $\bigcup_{x \in F} B(x,\varepsilon) \subset N$

Convex Sets in $\mathbb{R}^n$

A convex combination of a finite set of vectors $\{x_1, \dots, x_m\}$ in $\mathbb{R}$ is any vector $x = \sum_{i=1}^{m} \alpha_i x_i$ where $\sum_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$ for $i = 1..m$

The convex hull of $\{x_1, \dots, x_m\}$ is the set of all convex comb.s of $\{x_1, \dots, x_m\}$

A cone is a set F s.t. for all $x \in F$, we have that $x \in F \Rightarrow \alpha x \in F$ $\forall \alpha > 0$

Cone generated by $\{x_1 \cdots x_m\}$ is set of all $x = \sum_{i=1}^{m} \alpha_i x_i \quad \alpha_i \geq 0 \ (i=1..m)$

(all such cones convex)

Affine set in $\mathbb{R}^n$ is $\{x \mid x = y + s$ where $y \in \mathbb{R}^n$ and $s \in S$ and S subspace of $\mathbb{R}^n\}$

Affine hull of F (aff F) is smallest affine set containing F

Relative interior of F is int F relative to aff F. If $x \in F$ then $x \in \text{ri} F$ if $\varepsilon > 0$ exists s.t. $B(x, \varepsilon) \cap \text{aff} F \subset F$

(see examples in book)

Continuity and Limits

$f: D \subset \mathbb{R}^n \to \mathbb{R}^m$, for $x_0 \in \text{cl} D$ we say/write $\lim_{x \to x_0} f(x) = f_0$ if for all $\varepsilon > 0$

there exists $\delta > 0$ s.t.

$\|x - x_0\| < \delta$ and $x \in D \Rightarrow \|f(x) - f_0\| < \varepsilon$.

f continuous at x_0 if $x_0 \in D$ and $f_0 = f(x_0)$

In $\mathbb{R}$ we can define one-sided limits

$x \downarrow x_0 \quad (x_0 < x < x_0 + \delta)$

$x \uparrow x_0 \quad (x_0 - \delta < x < x_0)$

In general we can define one-sided limits on (along) a line

Let $f: D \to \mathbb{R}^m$ where $D \subset \mathbb{R}^n$

f is said to be Lipschitz continuous on some set $N \subset D$ if a constant $L > 0$ exists s.t.

$\|f(x_1) - f(x_0)\| \leq L\|x_1 - x_0\|$ for all $x_0, x_1 \in N$

f locally Lip. cont. at $\bar{x} \in \text{int} D$ if there is a n.hood N of $\bar{x}$ with $N \subset D$ s.t.

$\|f(x_1) - f(x_0)\| \leq L\|x_1 - x_0\|$ for all $x_1, x_0 \in N$

Sum of Lip. cont functions is Lip. cont

Product of Lip. cont functions $\mathbb{R}^n \to \mathbb{R}$ that are both bounded on N is Lip. cont. on N.

Consider $f: \mathbb{R}^n \to \mathbb{R}$

f diff. at x if $g \in \mathbb{R}^n$ exists s.t.

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - g^T y}{\|y\|} = 0$$

$\|y\|$ any norm (indep of path $y \to 0$)

(Frechet differentiable)

Directional derivative in direction p at x

$$D(f(x); p) = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon p) - f(x)}{\varepsilon}$$

Given x and p and $J = \left(\partial r_j / \partial x_i \right)$

defined and cont.

$$r(x+p) - r(x) = \int_0^1 J(x+\alpha p)\, p \, d\alpha$$

$\rightarrow$ for small p $r(x+p) \simeq r(x) + J(x)p$

If J lip cont in (connected) neighborhood containing

x and $x+p$ with lip constant L

$$\|r(x+p) - r(x) - J(x)p\| =$$

$$\left\| \int_0^1 [J(x+\alpha p) - J(x)]\, p \, d\alpha \right\| \leq$$

$$\int_0^1 \|J(x+\alpha p) - J(x)\| \, \|p\| \, d\alpha \leq$$

$$\int_0^1 L\alpha \|p\|^2 d\alpha = \tfrac{1}{2} L \|p\|^2$$