Methods for Unconstrained Optimization
Numerical Optimization Lectures 1-2

Coralia Cartis, University of Oxford

INFOMM CDT: Modelling, Analysis and Computation of
Continuous Real-World Problems
Problems and solutions

minimize $f(x)$ subject to $x \in \Omega \subseteq \mathbb{R}^n$.  

- $f : \Omega \rightarrow \mathbb{R}$ is (sufficiently) smooth ($f \in C^i(\Omega), i \in \{1, 2\}$).
- $f$ objective; $x$ variables; $\Omega$ feasible set (determined by finitely many constraints).
- $n$ may be large.
- minimizing $-f(x) \equiv -$ maximizing $f(x)$. Wlog, minimize.

$x^*$ global minimizer of $f$ over $\Omega \iff f(x) \geq f(x^*), \forall x \in \Omega$.

$x^*$ local minimizer of $f$ over $\Omega \iff$ there exists $\mathcal{N}(x^*, \delta)$ such that $f(x) \geq f(x^*)$, for all $x \in \Omega \cap \mathcal{N}(x^*, \delta)$,
where $\mathcal{N}(x^*, \delta) := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$ and $\|\cdot\|$ is the Euclidean norm.
Example problem in one dimension

Example: \( \min f(x) \) subject to \( a \leq x \leq b \).

- The feasible region \( \Omega \) is the interval \([a, b]\).
- The point \( x_1 \) is the global minimizer; \( x_2 \) is a local (non-global) minimizer; \( x = a \) is a constrained local minimizer.
Example problems in two dimensions

Ackley’s test function

Rosenbrock’s test function

[see Wikipedia]
Main classes of continuous optimization problems

Linear (Quadratic) programming: linear (quadratic) objective and linear constraints in the variables

\[
\min_{x \in \mathbb{R}^n} c^T x \left( + \frac{1}{2} x^T H x \right) \text{ subject to } a_i^T x = b_i, i \in E; \ a_i^T x \geq b_i, i \in I,
\]

where \( c, a_i \in \mathbb{R}^n \) for all \( i \) and \( H \) is \( n \times n \) matrix; \( E \) and \( I \) are finite index sets.

Unconstrained (Constrained) nonlinear programming

\[
\min_{x \in \mathbb{R}^n} f(x) \ (\text{subject to } c_i(x) = 0, i \in E; \ c_i(x) \geq 0, i \in I)
\]

where \( f, c_i : \mathbb{R}^n \to \mathbb{R} \) are (smooth, possibly nonlinear) functions for all \( i \); \( E \) and \( I \) are finite index sets.
Most real-life problems are nonlinear, often large-scale!
Example: an OR application

Optimization of a high-pressure gas network

pressures $p = (p_i, i)$; flows $q = (q_j, j)$; demands $d = (d_k, k)$; compressors. Maximize net flow s.t. the constraints:

$$\begin{align*}
A q - d &= 0 \\
A^T p^2 + K q^{2.8359} &= 0 \\
A^T_2 q + z \cdot c(p, q) &= 0 \\
\min p &\leq p \leq \max p \\
\min q &\leq q \leq \max q
\end{align*}$$

- $A, A_2 \in \{\pm 1, 0\}$; $z \in \{0, 1\}$
- 200 nodes and pipes, 26 machines: 400 variables;
- variable demand, $(p, d)$ 10mins. → 58,000 vars; real-time.
Example: an inverse problem application

Data assimilation for weather forecasting

- best estimate of the current state of the atmosphere
  - → find initial conditions $x_0$ for the numerical forecast
  by solving the (ill-posed) nonlinear inverse problem

$$
\min_{x_0} \sum_{i=0}^{m} (H_i[x_i] - y_i)^T R_i^{-1} (H[x_i] - y_i),
$$

$x_i = S(t_i, t_0, x_0)$, $S$ solution operator of the discrete nonlinear model; $H_i$ maps $x_i$ to observations $y_i$, $R_i$ error covariance matrix of the observations at $t_i$

$x_0$ of size $10^7 - 10^8$;
observations $m \approx 250,000$. 
Optimality conditions for unconstrained problems

== algebraic characterizations of solutions → suitable for computations.

- provide a way to guarantee that a candidate point is optimal (sufficient conditions)
- indicate when a point is not optimal (necessary conditions)

\[
\text{minimize } f(x) \text{ subject to } x \in \mathbb{R}^n. \quad \text{(UP)}
\]

First-order necessary conditions: \( f \in C^1(\mathbb{R}^n); \)
\( x^* \) a local minimizer of \( f \) \( \implies \nabla f(x^*) = 0. \)

\( \nabla f(x) = 0 \iff x \text{ stationary point of } f. \)
Lemma. Let $f \in C^1$, $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^n$ with $s \neq 0$. Then
\[ \nabla f(x)^T s < 0 \quad \implies \quad f(x + \alpha s) < f(x), \quad \forall \alpha > 0 \text{ sufficiently small.} \]
Proof. $f \in C^1 \implies \exists \alpha > 0$ such that
\[ \nabla f(x + \alpha s)^T s < 0, \quad \forall \alpha \in [0, \alpha]. \quad (\diamond) \]
Taylor’s/Mean value theorem:
\[ f(x + \alpha s) = f(x) + \alpha \nabla f(x + \tilde{\alpha} s)^T s, \quad \text{for some } \tilde{\alpha} \in (0, \alpha). \]
(\diamond) $\implies$ $f(x + \alpha s) < f(x), \quad \forall \alpha \in [0, \alpha]. \quad \square$

• $s$ descent direction for $f$ at $x$ if $\nabla f(x)^T s < 0$.

Proof of 1st order necessary conditions. assume $\nabla f(x^*) \neq 0$. $s := -\nabla f(x^*)$ is a descent direction for $f$ at $x = x^*$:
\[ \nabla f(x^*)^T (-\nabla f(x^*)) = -\nabla f(x^*)^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0 \]
since $\nabla f(x^*) \neq 0$ and $\|a\| \geq 0$ with equality iff $a = 0$. Thus, by Lemma, $x^*$ is not a local minimizer of $f$. \quad \square
Optimality conditions for unconstrained problems...

- $-\nabla f(x)$ is a descent direction for $f$ at $x$ whenever $\nabla f(x) \neq 0$.
- $s$ descent direction for $f$ at $x$ if $\nabla f(x)^T s < 0$, which is equivalent to

$$\cos\langle -\nabla f(x), s \rangle = \frac{(-\nabla f(x))^T s}{\|\nabla f(x)\| \cdot \|s\|} = \frac{|\nabla f(x)^T s|}{\|\nabla f(x)\| \cdot \|s\|} > 0,$$

and so:

$$\langle -\nabla f(x), s \rangle \in [0, \pi/2).$$

A descent direction $p_k$. 
Summary of first-order conditions. A look ahead

minimize $f(x)$ subject to $x \in \mathbb{R}^n$. (UP)

First-order necessary optimality conditions: $f \in C^1(\mathbb{R}^n)$; $x^*$ a local minimizer of $f \implies \nabla f(x^*) = 0$. 

$\tilde{x} = \arg \max_{x \in \mathbb{R}^n} f(x) \\
\quad \downarrow \\
\nabla f(\tilde{x}) = 0.$

- Look at higher-order derivatives to distinguish between minimizers and maximizers.
- ... except for convex functions: $x^*$ local minimizer/stationary point $\implies x^*$ global minimizer.

[see Problem Sheet]
Second-order optimality conditions (nonconvex fcts.)

Second-order necessary conditions: \( f \in C^2(\mathbb{R}^n); \)
\( x^* \) local minimizer of \( f \) \( \implies \) \( \nabla^2 f(x^*) \) positive semidefinite, i.e.,
\( s^T \nabla^2 f(x^*) s \geq 0, \) for all \( s \in \mathbb{R}^n. \)

Example: \( f(x) := x^3, \) \( x^* = 0 \) not a local minimizer but \( f'(0) = f''(0) = 0. \)

Second-order sufficient conditions: \( f \in C^2(\mathbb{R}^n); \)
\( \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \) positive definite, namely,
\( s^T \nabla^2 f(x^*) s > 0, \) for all \( s \neq 0. \)

\( \implies x^* \) (strict) local minimizer of \( f. \)

Example: \( f(x) := x^4, \) \( x^* = 0 \) is a (strict) local minimizer but \( f''(0) = 0. \)
Stationary points of quadratic functions

(a) Minimum  (b) Maximum  (c) Saddle  (d) Semidefinite

(e) Maximum or minimum

\[ H \in \mathbb{R}^{n \times n}, \ g \in \mathbb{R}^n: \quad q(x) := g^T x + \frac{1}{2} x^T H x. \]
\[ \nabla q(x^*) = 0 \iff H x^* + g = 0; \quad \nabla^2 q(x) = H, \ \forall x. \]

General \( f \): approx. locally quadratic around \( x^* \) stationary.
minimize $f(x)$ subject to $x \in \mathbb{R}^n$ (UP) \quad [f \in C^1(\mathbb{R}^n) \text{ or } f \in C^2(\mathbb{R}^n)]$

**A Generic Method (GM)**

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.

While (TERMINATION CRITERIA not achieved), REPEAT:

- compute the change
  \[x^{k+1} - x^k = F(x^k, \text{problem data}), \quad \text{[linesearch, trust-region]}\]

  to ensure $f(x^{k+1}) < f(x^k)$.
- set $x^{k+1} := x^k + F(x^k, \text{prob. data}), \quad k := k + 1$. \□

- **TC:** $\|\nabla f(x^k)\| \leq \epsilon$; maybe also, $\lambda_{\min}(\nabla^2 f(x^k)) \geq -\epsilon$.
- e.g., $x^{k+1} \equiv$ minimizer of some (simple) model of $f$ around $x^k$ \quad \longrightarrow \quad \text{linesearch, trust-region methods.}$
- if $F = F(x_k, x_{k-1}, \text{problem data}) \quad \longrightarrow \quad \text{conjugate gradients mthd.}$
Issues to consider about GM

**Global convergence of GM:**
if \( \epsilon := 0 \) and any \( x^0 \in \mathbb{R}^n: \nabla f(x^k) \rightarrow 0, \) as \( k \rightarrow \infty \) ?

[maybe also, \( \liminf_{k \rightarrow \infty} \lambda_{\min}(\nabla^2 f(x^k)) \geq 0 \) ?]

**Local convergence of GM:**
if \( \epsilon := 0 \) and \( x^0 \) sufficiently close to \( x^* \equiv \) stationary/local minimizer of \( f: x^k \rightarrow x^*, k \rightarrow \infty \) ?

**Global/local complexity of GM:** count number of iterations and their cost required by GM to generate \( x^k \) within desired accuracy \( \epsilon > 0 \), e.g., such that \( \|\nabla f(x^k)\| \leq \epsilon \).

[connection to convergence and its rate]

Rate of global/local convergence of GM.
Rates of convergence of sequences: an example

\[ l^k := (1/2)^k \longrightarrow 0 \] linearly,
\[ q^k := (1/2)^{2^k} \longrightarrow 0 \] quadratically,
\[ s^k := k^{-k} \longrightarrow 0 \] superlinearly as \( k \longrightarrow \infty \).

Rates of convergence on a log scale.

Notation: \((-i) := 10^{-i}\).
Rates of convergence of sequences

\( \{x^k\} \subset \mathbb{R}^n, x^* \in \mathbb{R}^n; \quad x^k \to x^* \text{ as } k \to \infty. \)

**p-Rate of convergence:** \( x^k \to x^* \) with rate \( p \geq 1 \) if \( \exists \rho > 0 \) and \( k_0 \geq 0 \) such that

\[
\|x^{k+1} - x^*\| \leq \rho \|x^k - x^*\|^p, \quad \forall k \geq k_0.
\]

- \( \rho \) convergence factor; \( e^k := x^k - x^* \) error in \( x^k \approx x^* \).

**Linear convergence:** \( p = 1 \quad \Rightarrow \quad \rho < 1 \); (asymptotically,) no of correct digits grows linearly in the number of iterations.

**Quadratic convergence:** \( p = 2 \); (asymptotically,) no of correct digits grows exponentially in the number of iterations.

**Superlinear convergence:** \( \|x^{k+1} - x^*\|/\|x^k - x^*\| \to 0 \) as \( k \to \infty \). [faster than linear, slower than quadratic; practically very acceptable]
Linesearch methods
A generic linesearch method

(UP): minimize $f(x)$ subject to $x \in \mathbb{R}^n$, where $f \in C^1$ or $C^2$.

**A Generic Linesearch Method (GLM)**

Choose $\epsilon > 0$ and $x^0 \in \mathbb{R}^n$.
While $\|\nabla f(x^k)\| > \epsilon$, REPEAT:

- compute a **descent** search direction $s^k \in \mathbb{R}^n$, 

\[ \nabla f(x^k)^T s^k < 0; \]

- compute a stepsize $\alpha^k > 0$ along $s^k$ such that 

\[ f(x^k + \alpha^k s^k) < f(x^k); \]

- set $x^{k+1} := x^k + \alpha^k s^k$ and $k := k + 1$.  

Recall property of descent directions.
Performing a linesearch

How to compute $\alpha^k$?

Exact linesearch:

$$\alpha^k := \arg \min_{\alpha > 0} f(x^k + \alpha s^k).$$

- computationally expensive for nonlinear objectives.

Exact linesearch for quadratic functions
Performing a linesearch ...

Inexact linesearch

- want stepsize $\alpha^k$ not to be:
  "too short"

- "too long" compared to the decrease in $f$
Performing a linesearch ...

Inexact linesearch

- want stepsizes $\alpha^k$ not “too long” compared to the decrease in $f$.

The Armijo condition

Choose $\beta \in (0, 1)$.

Compute $\alpha^k > 0$ such that

$$f(x^k + \alpha^k s^k) \leq f(x^k) + \beta \alpha^k \nabla f(x^k)^T s^k \quad (\ast)$$

is satisfied. □

- in practice, $\beta := 0.1$ or even $\beta := 0.001$.

- due to the descent condition, $\exists \overline{\alpha}^k > 0$ such that $(\ast)$ holds for $\alpha \in [0, \overline{\alpha}^k]$. Choose $\alpha^k \in (0, \overline{\alpha}^k]$ as large as possible.
Performing a linesearch ...

Inexact linesearch ...

- \( \Phi_k : \mathbb{R} \rightarrow \mathbb{R} \), \( \Phi_k(\alpha) := f(x^k + \alpha s^k) \), \( \alpha \geq 0 \). Then

\[
\text{Armijo} \iff \Phi_k(\alpha^k) \leq \Phi_k(0) + \beta \alpha^k \Phi'(0).
\]

Let \( y_\beta(\alpha) := \Phi_k(0) + \beta \alpha \Phi'(0) \), \( \alpha \geq 0 \).
Performing a linesearch ...

Inexact linesearch

The backtracking-Armijo (bArmijo) linesearch algorithm

Choose $\alpha(0) > 0$, $\tau \in (0, 1)$ and $\beta \in (0, 1)$.

While $f(x^k + \alpha(i)s^k) > f(x^k) + \beta \alpha(i) \nabla f(x^k)^T s^k$, REPEAT:

- set $\alpha(i+1) := \tau \alpha(i)$ and $i := i + 1$.

END.

Set $\alpha^k := \alpha(i)$. □

- backtracking $\implies$ stepsize $\alpha^k$ not “too short”;

$\alpha(0) := 1; \tau := 0.5 \implies \alpha(0) := 1, \alpha(1) := 0.5, \alpha(2) := 0.25, ...$

- the bArmijo linesearch algorithm terminates in a finite number of steps with $\alpha^k > 0$, due to the descent condition.
Steepest descent method

Steepest descent (SD) direction: set $s^k := -\nabla f(x^k)$, $k \geq 0$.

- $s^k$ descent direction whenever $\nabla f(x^k) \neq 0$:
  $$\nabla f(x^k)^T s^k < 0 \iff \nabla f(x^k)^T (-\nabla f(x^k)) < 0 \iff -\|\nabla f(x^k)\|^2 < 0.$$  

- $s^k$ steepest descent: unique global solution of
  $$\text{minimize}_{s \in \mathbb{R}^n} f(x^k) + s^T \nabla f(x^k) \quad \text{subject to} \quad \|s\| = \|\nabla f(x^k)\|.$$  

Cauchy-Schwarz: $|s^T \nabla f(x^k)| \leq \|s\| \cdot \|\nabla f(x^k)\|$, $\forall s$, with equality iff $s$ is proportional to $\|\nabla f(x^k)\|$.

Method of steepest descent (SD): GLM with $s^k = SD$ direction; any linesearch.

- SD-e := SD method with exact linesearches;
- SD-bA := SD method with bArmijo linesearches.
Global convergence of steepest descent methods

- $f \in C^1(\mathbb{R}^n)$; $\nabla f$ is Lipschitz continuous (on $\mathbb{R}^n$) iff $\exists L > 0,$ $ \| \nabla f(y) - \nabla f(x) \| \leq L \| y - x \|, \ \forall x, y \in \mathbb{R}^n.$

**Theorem** Let $f \in C^1(\mathbb{R}^n)$ be bounded below on $\mathbb{R}^n.$ Let $\nabla f$ be Lipschitz continuous. Apply the SD-e or the SD-bA method to minimizing $f$ with $\epsilon := 0.$

Then both variants of the SD method have the property:

either

- there exists $l \geq 0$ such that $\nabla f(x^l) = 0$

or

- $\| \nabla f(x^k) \| \to 0$ as $k \to \infty.$ $\square$

SD methods have excellent global convergence properties (under weak assumptions).
Some disadvantages of steepest descent methods

- SD methods are **scale-dependent**.

  poorly scaled variables $\implies$ SD direction gives little progress.

---

The effect of problem scaling on SD-e performance.

Let $f(x_1, x_2) = \frac{1}{2}(ax_1^2 + x_2^2)$, where $a > 0$. Left figure: $a = 10^{0.6}$ (mildly poor scaling). Right figure: $a = 1$ (“perfect” scaling).
Some disadvantages of steepest descent methods ...

- Usually, SD methods converge very slowly to solution, asymptotically.

theory: very slow converge.

numerics: break-down (cumulation of round-off and ill-conditioning).

\[ f(x_1, x_2) = 10(x_2 - x_1^2)^2 + \ldots + (x_1 - 1)^2. \]

SD-bA applied to the Rosenbrock function \( f \).
Local rate of convergence for steepest descent

Locally, SD converges linearly to a solution, namely

$$|f(x^{k+1}) - f(x^*)| \leq \rho |f(x^k) - f(x^*)|, \forall k \text{ suff. large}$$

BUT convergence factor $\rho$ v. close to 1 usually!

**Theorem.** $f \in C^2$; $x^*$ local minimizer of $f$ with $\nabla^2 f(x^*)$ positive definite $\rightarrow \lambda_{\text{max}}, \lambda_{\text{min}}$ eigenvalues.

Apply SD-e to $\min f$. If $x^k \rightarrow x^*$ as $k \rightarrow \infty$, then $f(x^k)$ converges linearly to $f(x^*)$, with

$$\rho \leq \left( \frac{\kappa(x^*) - 1}{\kappa(x^*) + 1} \right)^2 := \rho_{SD},$$

where $\kappa(x^*) = \lambda_{\text{max}}/\lambda_{\text{min}}$ condition number of $\nabla^2 f(x^*)$.

- **practice:** $\rho = \rho_{SD}$;
  
  for Rosenbrock $f$: $\kappa(x^*) = 258.10$, $\rho_{SD} \approx 0.984$.

- $\kappa(x^*) = 800$, $f(x^0) = 1$, $f(x^*) = 0$. SD-e gives $f(x^k) \approx 0.007$ after 1000 iterations!
Other directions for GLMs

Let $B^k$ symmetric, positive definite matrix. Let $s^k$ be def. by

$$B^k s^k = -\nabla f(x^k).$$

\[(\star)\]

$\implies$ $s^k$ descent direction;

$\implies$ $s^k$ solves the problem

$$\minimize_{s \in \mathbb{R}^n} m_k(s) = f(x^k) + \nabla f(x^k)^T s + \frac{1}{2} s^T B^k s^k.$$

If $\nabla^2 f(x^k)$ is positive definite and $B^k := \nabla^2 f(x^k)$

$\implies$ damped Newton method.

\[(\star)\] is a scaled steepest descent direction;

resulting GLMs can be made scale-invariant for some $B^k$
(e.g., $B^k = \nabla^2 f(x^k)$);

local convergence can be made faster than SD methods for
some $B^k$ (e.g., quadratic for Newton);
Global convergence for general GLMs

Theorem  Let \( f \in C^1(\mathbb{R}^n) \) be bounded below on \( \mathbb{R}^n \).
Let \( \nabla f \) be Lipschitz continuous. Let the eigenvalues of \( B^k \) be uniformly bounded above and below, away from zero, for all \( k \).
Apply the GLM to \( \min f \) with bArmijo linesearch, \( s^k \) from (\( \ast \)) and \( \epsilon = 0 \). Then

either

there exists \( l \geq 0 \) such that \( \nabla f(x^l) = 0 \)

or

\[ \| \nabla f(x^k) \| \to 0 \text{ as } k \to \infty. \]

\( \square \)

● Theorem requires locally strongly convex quadratic models of \( f \) for all \( k \). In particular, Th. holds if \( f \) locally strongly convex at all iterates.
● See Problem Sheet for failure examples for Newton’s method (without linesearch) to converge globally.
Local convergence for damped Newton’s method

Theorem  Let $f \in C^2(\mathbb{R}^n)$ and $\nabla^2 f$ be Lipschitz continuous and uniformly positive definite.

Let $s^k = \text{Newton direction in GLM}$ and let bArmijo linesearch have $\beta \leq 0.5$ and $\alpha(0) = 1$.

Then, provided limit points of the sequence of iterates exist,

- $\alpha^k = 1$ for all sufficiently large $k$;
- $x^k \rightarrow x^*$ as $k \rightarrow \infty$ quadratically.  □
Local convergence for damped Newton...

\[ f(x_1, x_2) = 10(x_2 - x_1^2)^2 + (x_1 - 1)^2; \quad x^* = (1, 1). \]

Damped Newton with bArmijo linesearch applied to the Rosenbrock function \( f \).
Modified Newton methods

If $\nabla^2 f(x^k)$ is not positive definite, it is usual to solve instead

$$(\nabla^2 f(x^k) + M^k) s^k = -\nabla f(x^k),$$

where

- $M^k$ chosen such that $\nabla^2 f(x^k) + M^k$ is “sufficiently” positive definite.
- $M^k := 0$ when $\nabla^2 f(x^k)$ is “sufficiently” positive definite.

**Popular option:** Modified Cholesky: compute Cholesky factorization

$$\nabla^2 f(x^k) = L^k (L^k)^\top$$

where $L^k$ is lower triangular matrix. Modify the generated $L^k$ if the factorization is in danger of failing (modify small or negative diagonal pivots, etc.).
Quasi-Newton methods

**Secant approximations** for computing $B^k \approx \nabla^2 f(x^k)$

[when $\nabla^2 f$ unavailable exactly or by function or gradient finite-differences]

At the start of the GLM, choose $B^0$ (say, $B^0 := I$). After computing $s^k$ from $B^k s^k = -\nabla f(x^k)$ and $x^{k+1} = x^k + \alpha^k s^k$, compute update $B^{k+1}$ of $B^k$.

**Wish list:**

Compute $B^{k+1}$ as a function of already-computed quantities $\nabla f(x^{k+1}), \nabla f(x^k), \ldots, \nabla f(x^0), B^k, s^k$

$B^{k+1}$ should be symmetric, nonsingular (pos. def?)

$B^{k+1}$ ‘close’ to $B^k$, a ‘cheap’ update of $B^k$

$B^k \rightarrow \nabla^2 f(x^k)$, etc.

⇒ a new class of methods: faster than steepest descent method, cheaper to compute per iteration than Newton’s.
Quasi-Newton methods ...

For the first wish, choose $B^{k+1}$ to satisfy the secant equation

$$\gamma^k := \nabla f(x^{k+1}) - \nabla f(x^k) = B^{k+1}(x^{k+1} - x^k) = B^{k+1}\alpha^k s^k.$$  

- It is satisfied by $B^{k+1} = \nabla^2 f$ when $f$ is a quadratic function.
- The change in gradient contains information about the Hessian.

Many ways to compute $B^{k+1}$ to satisfy secant eq; trade-off between wishes: rank-1 update of $B^k$ (Symmetric-Rank 1), rank 2 update of $B^k$ (BFGS (Broyden-Fletcher-Goldfarb-Shanno)), DFP (Davidson-Fletcher-Powell), etc.

- Work per iteration: $\mathcal{O}(n^2)$ (as opposed to the $\mathcal{O}(n^3)$ of Newton)
- For global convergence, must use Wolfe linesearch instead of bArmijo linesearch.
- **local Q-superlinear convergence**

[see Problem Sheet]
Trust-region methods
Linesearch versus trust-region methods

(UP): minimize \( f(x) \) subject to \( x \in \mathbb{R}^n \).

**Linesearch methods:** “liberal” in the choice of search direction, keeping bad behaviour in control by choice of \( \alpha^k \).

- choose descent direction \( s^k \),
- compute stepsize \( \alpha^k \) to reduce \( f(x^k + \alpha s^k) \),
- update \( x^{k+1} := x^k + \alpha^k s^k \).

**Trust region (TR) methods:** “conservative” in the choice of search direction, so that a full stepsize along it may really reduce the objective.

- pick direction \( s^k \) to reduce a “local model” of \( f(x^k + s^k) \),
- accept \( x^{k+1} := x^k + s^k \) if decrease in the model is also achieved by \( f(x^k + s^k) \),
- else set \( x^{k+1} := x^k \) and “refine” the model.
Trust-region models for unconstrained problems

Approximate $f(x^k + s)$ by:

- **linear model**
  \[ l_k(s) := f(x^k) + s^\top \nabla f(x^k) \]

- **quadratic model**
  \[ q_k(s) := f(x^k) + s^\top \nabla f(x^k) + \frac{1}{2} s^\top \nabla^2 f(x^k) s. \]

**Impediments:**

- models may not resemble $f(x^k + s)$ when $s$ is large,
- models may be unbounded from below,
  - $l_k(s)$ **always** unbounded below (unless $\nabla f(x^k) = 0$)
  - $q_k(s)$ **is always** unbounded below if $\nabla^2 f(x^k)$ is negative definite or indefinite, and **sometimes** if $\nabla^2 f(x^k)$ is positive semidefinite.
Trust region models and subproblem

Prevent bad approximations by trusting the model only in a trust region, defined by the trust region constraint

$$\|s\| \leq \Delta_k,$$  \hspace{1cm} (R)

for some “appropriate” radius $\Delta_k > 0$.

The constraint (R) also prevents $l_k, q_k$ from unboundedness!

$\implies$ the trust region subproblem

$$\min_{s \in \mathbb{R}^n} m_k(s) \text{ subject to } \|s\| \leq \Delta_k,$$  \hspace{1cm} (TR)

where $m_k := l_k, k \geq 0$, or $m_k := q_k, k \geq 0$.

- From now on, $m_k := q_k$.

(TR) easier to solve than (P). May even solve (TR) only approximately.
Trust region models and subproblem - an example

Trust-region models of $f(x) = x_1^4 + x_1 x_2 + (1 + x_2)^2$. 
Generic trust-region method

Let $s^k$ be a(n approximate) solution of (TR). Then

- **predicted model decrease:**
  $$m_k(0) - m_k(s^k) = f(x^k) - m_k(s^k).$$

- **actual function decrease:**
  $$f(x^k) - f(x^k + s^k).$$

The trust region radius $\Delta_k$ is chosen based on the value of

$$\rho_k := \frac{f(x^k) - f(x^k + s^k)}{f(x^k) - m_k(s^k)}.$$

If $\rho_k$ is not too smaller than 1, $x^{k+1} := x^k + s^k$, $\Delta_{k+1} \geq \Delta_k$.

If $\rho_k$ close to or $\geq 1$, $\Delta_k$ is increased.

If $\rho_k \ll 1$, $x^{k+1} = x^k$ and $\Delta_k$ is reduced.
A Generic Trust Region (GTR) method

Given $\Delta_0 > 0$, $x^0 \in \mathbb{R}^n$, $\epsilon > 0$. While $\|\nabla f(x^k)\| \geq \epsilon$, do:

1. Form the local quadratic model $m_k(s)$ of $f(x^k + s)$.

2. Solve (approximately) the (TR) subproblem for $s^k$ with $m_k(s^k) < f(x^k)$ ("sufficiently").

Compute

$$\rho_k := \frac{f(x^k) - f(x^k + s^k)}{f(x^k) - m_k(s^k)}.$$

3. If $\rho_k \geq 0.9$, then [very successful step]

   set $x^{k+1} := x^k + s^k$ and $\Delta_{k+1} := 2\Delta_k$.

   Else if $\rho_k \geq 0.1$, then [successful step]

   set $x^{k+1} := x^k + s^k$ and $\Delta_{k+1} := \Delta_k$.

   Else [unsuccessful step]

   set $x^{k+1} = x^k$ and $\Delta_{k+1} := \frac{1}{2}\Delta_k$.

4. Let $k := k + 1$. 

□
Trust-region methods

• Other sensible values of the parameters of the GTR are possible.

“Solving” the (TR) subproblem

\[
\min_{s \in \mathbb{R}^n} m_k(s) \quad \text{subject to} \quad \|s\| \leq \Delta_k, \quad \text{(TR)}
\]

... exactly or even approximately may imply work.

Want “minimal” condition of “sufficient decrease” in the model that ensures global convergence of the GTR method (the Cauchy cond.). In practice, we (usually) do much better than this condition!

Example of applying a trust-region method: [Sartenaer, 2008].

■ approximate solution of (TR) subproblem: better than Cauchy, but not exact.

■ notation: \( \Delta f / \Delta m_k \equiv \rho_k \).
The Cauchy point of the (TR) subproblem

- recall the steepest descent method has strong (theoretical) global convergence properties; same will hold for GTR method with SD direction.

“minimal” condition of “sufficient decrease” in the model: require

$$m_k(s^k) \leq m_k(s_c^k) \text{ and } \|s^k\| \leq \Delta_k,$$

where $s_c^k := -\alpha_c^k \nabla f(x^k)$, with

$$\alpha_c^k := \arg \min_{\alpha > 0} m_k(-\alpha \nabla f(x^k)) \text{ subject to } \|\alpha \nabla f(x^k)\| \leq \Delta_k.$$

[i.e. a linesearch along steepest descent direction is applied to $m_k$ at $x^k$ and is restricted to the trust region.] Easy:

$$\alpha_c^k := \arg \min_{\alpha} m_k(-\alpha \nabla f(x^k)) \text{ subject to } 0 < \alpha \leq \frac{\Delta_k}{\|\nabla f(x^k)\|}.$$

- $y_c^k := x^k + s_c^k$ is the Cauchy point.  

[see Problem Sheet]
Global convergence of the GTR method

**Theorem** (GTR Global Convergence)

Let $f$ be smooth on $\mathbb{R}^n$ and bounded below on $\mathbb{R}^n$. Let $\nabla f$ be Lipschitz continuous on $\mathbb{R}^n$.

Let \( \{x^k\}, \; k \geq 0, \) be generated by the generic trust region (GTR) method, and let the computation of $s^k$ be such that $m_k(s^k) \leq m_k(s^k_c)$, for all $k$.

Then either

$$\text{there exists } k \geq 0 \text{ such that } \nabla f(x^k) = 0$$

or

$$\lim_{k \to \infty} \| \nabla f(x^k) \| = 0.$$
Solving the (TR) subproblem

On each TR iteration we compute or approximate the solution of

\[
\min_{s \in \mathbb{R}^n} m_k(s) = f(x^k) + s^\top \nabla f(x^k) + \frac{1}{2} s^\top \nabla^2 f(x^k)s
\]

subject to \( \|s\| \leq \Delta_k \).

also, \( s^k \) must satisfy the Cauchy condition \( m_k(s^k) \leq m_k(s^k_c) \), where \( s^k_c := -\alpha_c^k \nabla f(x^k) \), with

\[
\alpha_c^k := \arg \min_{\alpha > 0} m_k(-\alpha \nabla f(x^k)) \text{ subject to } \|\alpha \nabla f(x^k)\| \leq \Delta_k.
\]

[Cauchy condition ensures global convergence]

• solve (TR) exactly (i.e., compute global minimizer of TR) \( \implies \) TR akin to Newton-like method.

• solve (TR) approximately (i.e., an approximate global minimizer) \( \implies \) large-scale problems.
Solving the (TR) subproblem exactly

For $h \in \mathbb{R}$, $\Delta > 0$, $g \in \mathbb{R}^n$, $H \in \mathbb{R}^{n \times n}$ real matrix, consider

$$\min_{s \in \mathbb{R}^n} \ m(s) := h + s^\top g + \frac{1}{2} s^\top H s, \ \text{s. t.} \ \|s\| \leq \Delta.$$  \ (TR)

Characterization result for the solution of (TR):

**Theorem**

Any **global** minimizer $s^*$ of (TR) satisfies the equation

$$(H + \lambda^* I)s^* = -g,$$

where $H + \lambda^* I$ is positive semidefinite, $\lambda^* \geq 0$,

$$\lambda^*(\|s^*\| - \Delta) = 0 \quad \text{and} \quad \|s^*\| \leq \Delta.$$

If $H + \lambda^* I$ is positive definite, then $s^*$ is unique.

- The above Theorem gives necessary and sufficient **global** optimality conditions for a **nonconvex** optimization problem!
Solving the (TR) subproblem exactly

Computing the global solution \( s^* \) of (TR):

**Case 1.** If \( H \) is positive definite and \( Hs = -g \) satisfies \( \|s\| \leq \Delta \)

\[ \implies s^* := s \text{ (unique), } \lambda^* := 0 \text{ (by Theorem).} \]

**Case 2.** If \( H \) is positive definite but \( \|s\| > \Delta \),

or \( H \) is not positive definite, Theorem implies \( s^* \) satisfies

\[ (H + \lambda I)s = -g, \quad \|s\| = \Delta, \quad (\ast) \]

for some \( \lambda \geq \max\{0, -\lambda_{\min}(H)\} := \lambda \).

Let \( s(\lambda) = -(H + \lambda I)^{-1}g \), for any \( \lambda > \lambda \). Then \( s^* = s(\lambda^*) \)

where \( \lambda^* \geq \lambda \) solution of

\[ \|s(\lambda)\| = \Delta, \quad \lambda \geq \lambda. \]

\[ \implies \text{nonlinear equation in one variable } \lambda. \text{ Use Newton’s method to solve it.} \]
Further considerations on trust-region methods

- Solving the (TR) subproblem accurately for large-scale (and dense) problems uses iterative methods (conjugate-gradient, Lanczos); Cauchy condition satisfied

- Quasi-Newton methods/approximate derivatives also possible in the trust-region framework (no need for positive definite updates for the Hessian); replace $\nabla^2 f(x^k)$ with approximation $B^k$ in the quadratic local model $m_k(s)$.

**Conclusions:** state-of-the-art software exists implementing linesearch and TR methods; similar performances for linesearch and TR methods (more heuristical approaches needed by linesearch methods to deal with negative curvature). Choosing between the two is mostly a matter of “taste”.
Numerical optimization bibliography


Information on existing software can be found at the NEOS Center: [http://www.neos-guide.org](http://www.neos-guide.org)

→ look under **Optimization Guide** and **Optimization Tree**, etc.