# Nonlinear optimization

## Unconstrained nonlinear optimization - Lecture III

*François Glineur*

UCL/FSA/INMA & CORE

glineur@core.ucl.ac.be

Graduate School in Systems and Control

March 25 2004

# Administrivia

## Schedule

- Part I - *Convex optimization*: March 11 & 16
- Part II - *(Traditional) nonlinear optimization*: March 25 and April 1
- Part III - *Some applications in systems and control* (by Michael OVERTON): April 22 & 29

Coffee break at 10:45am ... with a poll!

## Questions and comments ...

... are more than welcome, at any time !

Slides should be available on the web :
http://www.core.ucl.ac.be/~glineur/

## References

Part II mainly based on a single recent reference

- ◇ Numerical Optimization, Jorge Nocedal and Stephen J. Wright, Springer, 1999

# Plan of Lecture III

## Two paradigms for unconstrained optimization

- Fundamentals of unconstrained optimization
  - What is a solution?
  - Two strategies
- Line search techniques
  - Step length selection and convergence
- Trust-region techniques
  - Model definition and convergence

# Unconstrained optimization

## Fundamentals

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

(Usually) assume $f$ is smooth

Reminder: universal algorithm does not exist!

Consider a *simple* unconstrained minimization

$$\min f(x_1, x_2, \ldots, x_{10})$$

with smooth $f$ (Lipschitz continuous with $L = 2$):

One can show that exists some functions where at least $10^{20}$ iterations (function evaluations) are needed to find a solution with accuracy 1% !

## What is a solution?

- ◇ Global minimizer $x^*$ iff $f(x^*) \leq f(x)\ \forall x$
  (but no hope of finding them)
- ◇ Local minimizer $x^*$ iff $f(x^*) \leq f(x)\ \forall x \in \mathcal{N}$
  with $\mathcal{N}$ some open neighborhood of $x^*$
- ◇ Strict local minimizer iff $f(x^*) < f(x)\ \forall x \neq x^* \in \mathcal{N}$
- ◇ Isolated local minimizer iff $x^*$ is the only strict minimizer in some neighborhood of $x^*$

We have strict inclusions

$$\text{Isolated} \Rightarrow \text{Strict} \Rightarrow \text{Local}$$

($x^4 \cos(1/x) + 2x^4$ has a strict min. in 0 but not isolated)

## Recognizing a local minimum

Main tools (assuming enough smoothness where necessary):

First order:

$$f(x+\Delta x) = f(x)+\nabla f(x+\alpha\Delta x)^{\mathrm{T}}\Delta x \text{ for some } 0 < \alpha < 1$$

and thus

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^{\mathrm{T}}\Delta x$$

Second order:

$$f(x+\Delta x) = f(x)+\nabla f(x)^{\mathrm{T}}\Delta x+\frac{1}{2}(\Delta x)^{\mathrm{T}}\nabla^2 f(x+\alpha\Delta x)\Delta x$$

for some $0 < \alpha < 1$ and thus

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^{\mathrm{T}}\Delta x + \frac{1}{2}(\Delta x)^{\mathrm{T}}\nabla^2 f(x)\Delta x$$

## Necessary and sufficient conditions

- $x^*$ local minimizer $\Rightarrow \nabla f(x^*) = 0$ (stationary point)
  ($\nabla f(x)$ continuous on neighborhood of $x^*$)
- $x^*$ local minimizer $\Rightarrow \nabla^2 f(x^*) \succeq 0$ (p.s.d.)
  ($\nabla^2 f(x)$ continuous on neighborhood of $x^*$)
- $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$ (p.d.)
  $\Rightarrow x^*$ strict local minimizer
  ($\nabla^2 f(x)$ continuous on neighborhood of $x^*$)
  But no sufficient condition for non-strict minimizer!
- $f$ convex $\Rightarrow$ local minimizer $\equiv$ global minimizer
  assuming $f$ differentiable, stationary $\equiv$ minimizer

## Two strategies

◇ Line search

– Choose direction $p_k$

– Choose step length $\alpha_k$ solving (approximately)

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

◇ Trust region

– Choose model $m_k$ such that

$$m_k(x_k + p_k) \approx f(x_k + p_k) \text{ around } x_k$$

– Choose trust region defined by $\|p_k\| \leq \Delta_k$

– Minimize model (approximately) over trust region

Somehow opposite strategies!

## Which line search direction?

- Descent direction when $\nabla f(x)^{\mathrm{T}} p < 0$
- What is the best descent direction ?

$$\min_p \nabla f(x)^{\mathrm{T}} p \text{ such that } \|p\| = 1$$

has solution

$$p^S = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

$\Rightarrow$ steepest descent direction

- Newton direction considering

$$f(x+p) \approx f(x) + \nabla f(x)^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x) p = 0$$

$$\Rightarrow p^N = -\nabla^2 f(x)^{-1} \nabla f(x) \text{ (assuming } \nabla^2 f(x) \text{ p.s.d.)}$$

## More about Newton direction

- $\nabla^2 f(x)$ p.s.d. $\Rightarrow$ $p^N$ is a descent direction
- Computing second derivatives is potentially expensive or error prone

  $\Rightarrow$ replace $\nabla^2 f(x)$ by approximation $B_k$

A sound requirement:

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_{k+1})(x_{k+1} - x_k) + o(\|x_{k+1} - x_k\|)$$

$$\Rightarrow \quad \nabla^2 f(x_{k+1})(x_{k+1} - x_k) \approx \nabla f(x_{k+1}) - \nabla f(x_k)$$

$$\Rightarrow \quad B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$$

$$\Rightarrow \quad B_{k+1} s_k = y_k$$

These are called quasi-Newton directions $-B_k^{-1} \nabla f(x_k)$

## Quasi-Newton directions

- Typically, impose symmetry on $B_k$ (mimic Hessian)
- Update $B_k$ with low-rank perturbation
  - Symmetric rank one (SR1)

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

  - BFGS (Broyden-Fletcher-Goldfarb-Shanno)

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

(rank two, $B_k$ p.d. if $B_0$ p.d. and $s_k^{\mathrm{T}} y_k > 0$)

- Equivalent formulae for $H_k = B_k^{-1} \Rightarrow p_k = -H_k \nabla f(x_k)$

## Which model for trust-region?

◇ Quadratic to ease minimization

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

Exact up to first order

◇ Case $B_k = 0$ ... (not useful)

$\Rightarrow$ steepest descent with step length depending on $\Delta_k$

◇ Case $B_k = \nabla^2 f(x_k)$

$\Rightarrow$ second-order model

◇ Case $B_k \approx \nabla^2 f(x_k)$ (e.g. SR1 or BFGS)

$\Rightarrow$ quasi-Newton trust region

Advantage: $\Delta_k \Rightarrow$ minimum exists (even when $B_k \not\succ 0$)

## Scaling issues

- ◇ Poor scaling can arise from model
$$f(x_1, x_2) = 10^{-4}x_1^3 - 10^5 x_2^2$$
- ◇ Choice of units
- ◇ Diagonal rescaling
$$\hat{x} = Dx \text{ with } D = \operatorname{diag} d_i > 0$$
- ◇ Some methods are sensitive to poor scaling
  (e.g. steepest descent)
  some others are not (e.g. Newton's method)

$\Rightarrow$ scale-invariance is a desirable property
(usually more difficult for TR than for LS)

# Line search

## Choosing the step length: Wolfe conditions

◇ *Sufficient decrease* condition (Armijo):

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^{\mathrm{T}} p_k$$

with $0 < c_1 < 1$ (typically $10^{-4}$)

Always possible to satisfy when $\alpha \to 0$

$\Rightarrow$ we also need ...

◇ *Curvature* condition

$$\nabla f(x_k + \alpha p_k)^{\mathrm{T}} p_k \geq c_2 \nabla f(x_k)^{\mathrm{T}} p_k$$

with $c_1 < c_2 < 1$ (typically 0.9 for a (quasi)-Newton)

## Strong Wolfe condition

Replace curvature condition by

$$\left|\nabla f(x_k + \alpha p_k)^{\mathrm{T}} p_k\right| \leq c_2 \left|\nabla f(x_k)^{\mathrm{T}} p_k\right|$$

## Meaning

Recall that

$$\phi(\alpha) = f(x_k + \alpha p_k) \Rightarrow \phi'(\alpha) = \nabla f(x + \alpha p_k)^{\mathrm{T}} p_k$$

- ◇ Sufficient decrease condition forces rate of decrease to be at least $c_1\phi'(0)$

$$\phi(\alpha) \leq \phi(0) - \alpha c_1 \phi'(0)$$

- ◇ Curvature condition bounds $\phi'(\alpha)$ (strong: $|\phi'(\alpha)|$)

$$\phi'(\alpha) \geq c_2 \phi'(0) \quad (\text{strong: } |\phi'(\alpha)| \leq c_2 \left|\phi'(0)\right|)$$

## Existence

Assume

- $p$ is a descent direction
- $\phi(\alpha) = f(x_k + \alpha p_k)$ is bounded below for $\alpha > 0$
- $0 < c_1 < c_2 < 1$

Then there are intervals of step lengths satisfying the Wolfe conditions and the strong Wolfe conditions

There exists a (one-dimensional) search procedure guaranteed to compute a point on this interval

These conditions are *scale-invariant*

## Backtracking

As an alternative to the second curvature condition:
Choose starting $\alpha > 0$ and $0 < \rho < 1$
(e.g. $\alpha = 1$ for (quasi-)Newton)

- While $f(x_k + \alpha p_k) > f(x_k) + c_1 \alpha \nabla f(x_k)^{\mathrm{T}} p_k$
- Update $\alpha$ with $\rho\alpha$

## In practice

Good $\alpha$s can be found by interpolation techniques using

- function values and
- derivatives previously computed

e.g. minimize cubic interpolant based on $\phi(0), \phi'(0), \phi(\alpha^{(i)})$ and $\phi(\alpha^{(i-1)})$ or on $\phi(\alpha^{(i)}), \phi(\alpha^{(i-1)}), \phi'(\alpha^{(i)})$ and $\phi'(\alpha^{(i-1)})$

## Convergence

Define angle $\theta_k$ between $p_k$ and $\nabla f(x_k)$ by

$$\cos\theta_k = -\frac{\nabla f(x_k)^{\mathrm{T}} p_k}{\|\nabla f(x_k)\| \, \|p_k\|}$$

Assuming $f$ bounded below, continuously differentiable, $p_k$ descent directions satisfying Wolfe conditions, $\nabla f$ is Lipschitz continuous, we have

$$\sum_{k\geq 0} \cos^2\theta_k \, \|\nabla f(x_k)\|^2 < +\infty \quad \text{(Zoutendijk condition)}$$

- Implies $\cos^2\theta_k \, \|\nabla f(x_k)\|^2 \to 0$
- If angle bounded away from $\frac{\pi}{2}$ i.e. $\cos\theta_k \geq \delta > 0$ then $\|\nabla f(x_k)\|^2 \to 0$ stationary pt (e.g. steepest descent)

## Convergence (continued)

- ◇ We only get stationary points
  since no second-order information is used
- ◇ (Quasi-)Newton:
  assuming
  $$\|B_k\| \leq M \text{ and } \left\|B_k^{-1}\right\| \leq M$$
  we have
  $$\cos\theta_k \geq \frac{1}{M}$$
  $\Rightarrow$ convergence when $B_k$
  - are p.d. (to ensure descent property) and
  - have bounded condition numbers

## Rate of convergence: steepest descent

◇ For a convex quadratic $f(x) = \frac{1}{2}x^{\mathrm{T}}Qx - b^{\mathrm{T}}x$
with exact line searches (and $\lambda_i$ eigenvalues of $Q \succ 0$)

$$\|x_{k+1} - x^*\|_Q \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \|x_k - x^*\|_Q$$

◇ In general with exact line searches and $\nabla^2 f(x^*) \succ 0$

$$f(x_{k+1}) - f(x^*) \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}(f(x_k) - f(x^*))$$

with $\lambda_i$ eigenvalues of $\nabla^2 f(x^*)$
$\rightarrow$ linear rate $\rightarrow$ slow (and inexact is worse)

$\kappa(Q) = 800, f(x_0) = 1, f(x^*) = 0 \Rightarrow f(x_{1000}) \approx 0.08$

## Rate of convergence: general descent

For a general descent direction $p_k$: if

- $\nabla^3 f$ continuous
- $x \to x^*$ such that
- $x^*$ minimizer with $\nabla f(x) = 0$ and $\nabla^2 f(x^*) \succ 0$
- $\alpha_k$ satisfies Wolfe with $c_1 \leq \frac{1}{2}$
- $\lim_{k\to\infty} \left\|\nabla f(x_k) + \nabla^2 f(x_k) p_k\right\| / \|p_k\| = 0$

Then

- $\alpha_k = 1$ becomes admissible for all $k \geq k_0$
- $x_k \to x^*$ superlinearly if $\alpha_k = 1$ is chosen $\forall\ k \geq k_0$

$\Rightarrow$ full step $\alpha_k^{(0)} = 1$ must be tried first

## Rate of convergence: quasi-Newton

For $p_k = B_k^{-1} \nabla f(x_k)$: if

- $\nabla^3 f$ continuous
- $x \to x^*$ such that
- $x^*$ minimizer with $\nabla f(x) = 0$ and $\nabla^2 f(x^*) \succ 0$
- $\alpha_k = 1 \; \forall k$

Then

- $x_k \to x^*$ superlinearly if and only if

$$\lim_{k \to \infty} \frac{\|(B_k - \nabla^2 f(x^*)) p_k\|}{\|p_k\|} = 0$$

- $B_k \to \nabla^2 f(x^*)$ not needed ! (only along $p_k$)

## Rate of convergence: Newton

For $p_k = \nabla^2 f(x_k)^{-1} \nabla f(x_k)$: if

- $\nabla^2 f$ Lipschitz continuous
- $x^*$ minimizer with $\nabla f(x) = 0$ and $\nabla^2 f(x^*) \succ 0$
- $x_0$ sufficiently close to $x^*$
- $\alpha_k = 1 \; \forall k$

Then

- $x_k \to x^*$
- quadratic rate of convergence (cf. previous slide)
- gradient norms $\|\nabla f(x_k)\|$ quadratically tend to 0

# Trust region

## Principle

◇ Model (to be trusted on region $\{\|x - x_k\| \leq \Delta_k\}$)

$$m_k(x_k + p_k) = f(x_k) + \nabla f(x_k)^{\mathrm{T}} p_k + \frac{1}{2} p_k^{\mathrm{T}} B_k p_k$$

exact up to first or second order

◇ Trust region radius $\Delta_k$:

– decrease when model is a bad approximation of $f$

– increase when model is a good approximation of $f$

◇ Actual criteria depends on $p_k$ according to

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(x_k) - m_k(x_k + p_k)} = \frac{\text{actual reduction}}{\text{predicted reduction}}$$

## Main algorithm

Given $\Delta_M$, $0 < \Delta_0 \leq \Delta_M$, $0 \leq \eta < \frac{1}{4}$
For $k = 0, 1, 2, \ldots$

- Obtain $p_k$ by solving (approximately)

$$\min m_k(x_k + p_k) \text{ such that } \|p_k\| \leq \Delta_k$$

- Compute $\rho_k$
- If $\rho_k < \frac{1}{4}$ set $\Delta_{k+1} = \frac{1}{4}\|p_k\|$
  If $\frac{1}{4} \leq \rho_k \leq \frac{3}{4}$ set $\Delta_{k+1} = \Delta_k$
  If $\frac{3}{4} < \rho_k$ set $\Delta_{k+1} = \min\{2\Delta_k, \Delta_M\}$
- If $\rho_k > \eta$ set $x_{k+1} = x_k + p_k$
  If $\rho_k \leq \eta$ set $x_{k+1} = x_k$

## Cauchy point

The Cauchy point is the model minimizer on the steepest descent direction

$$p_k^C = -\tau_k \Delta_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

with

$$\tau_k = 1$$

when $\nabla f(x_k)^{\mathrm{T}} B_k \nabla f(x_k) \leq 0$ or

$$\tau_k = \min\{1, \|\nabla f(x_k)\|^3 / (\Delta_k \nabla f(x_k)^{\mathrm{T}} B_k \nabla f(x_k))$$

when $\nabla f(x_k)^{\mathrm{T}} B_k \nabla f(x_k) > 0$

- Can be inside ($\tau_k < 1$) or on the boundary ($\tau_k = 1$)

## Convergence result

The Cauchy point achieves the following decrease

$$m_k(x_k) - m_k(x_k + p_k^C) \geq \frac{1}{2} \|\nabla f(x_k)\| \min\{\Delta_k, \frac{\|\nabla f(x_k)\|}{\|B_k\|}\}$$

If one can guarantee a reduction of the same order $\forall k$

$$m_k(x_k) - m_k(x_k + p_k) \geq c_1 \|\nabla f(x_k)\| \min\{\Delta_k, \frac{\|\nabla f(x_k)\|}{\|B_k\|}\}$$

assuming $\nabla f$ is continuous, $f$ is bounded below and a uniform bound $\|B_k\| \leq \beta \ \forall k$ we have

- When $\eta = 0$: $\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0$
- When $0 < \eta < \frac{1}{4}$: $\lim_{k \to \infty} \nabla f(x_k) = 0$

Only stationarity is guaranteed

## Strategies for computing a valid $p_k$

◇ Stick to $p_k^C$ (but second-order information not used)

◇ Dogleg: minimize on path $x_k \to x_k + p_k^U \to x_k + p_k^B$

$$p_k^U = -\frac{\nabla f(x_k)^{\mathrm{T}} \nabla f(x_k)}{\nabla f(x_k)^{\mathrm{T}} B_k \nabla f(x_k)} \nabla f(x_k)$$

(this is the minimum along $-\nabla f(x_k)$)

$$p^B = -B_k^{-1} \nabla f(x_k)$$

(this is actual model minimizer)

– path intersects trust region boundary at most once

– intersection can be computed easily (scalar quadratic)

but this approach requires that $B_k$ is pos. definite

◇ 2D subspace minimization: minimize on $x_k + \text{span}\{p_k^C, p_k^U\}$
(can be adapted when $B_k$ is not p.d.)

In all three cases (Cauchy, dogleg, 2D subspace):
Cauchy decrease condition satisfied $\Rightarrow$ global convergence

*Thanks for you attention*