Part 4

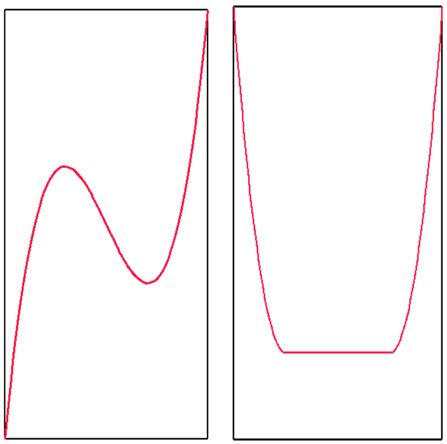
Smooth unconstrained problems:
Line search algorithms

$$\text{minimize} \quad f(x)$$

Smooth problems: Characterization of Optima

Problem: find solution x^* of

$$\text{minimize}_x \quad f(x)$$



A strict local minimum x^* must satisfy two conditions:

First order necessary condition: gradient must vanish:

$$\nabla f(x^*)=0$$

Sufficient condition for a strict minimum:

$$\text{spectrum}(\nabla^2 f(x^*)) > 0$$

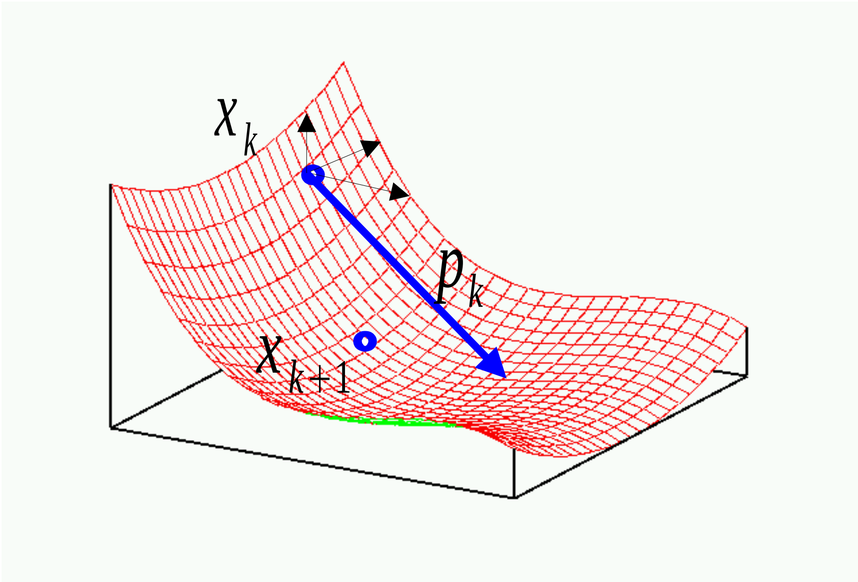
Basic Algorithm for Smooth Unconstrained Problems

Basic idea for iterative solution $x_k \to x^*$ of the problem

$$\text{minimize} \quad f(x)$$

Generate a sequence x_k by

1. finding a search direction p_k
2. choosing a step length α_k



Then compute the update

$$x_{k+1} = x_k + \alpha_k p_k$$

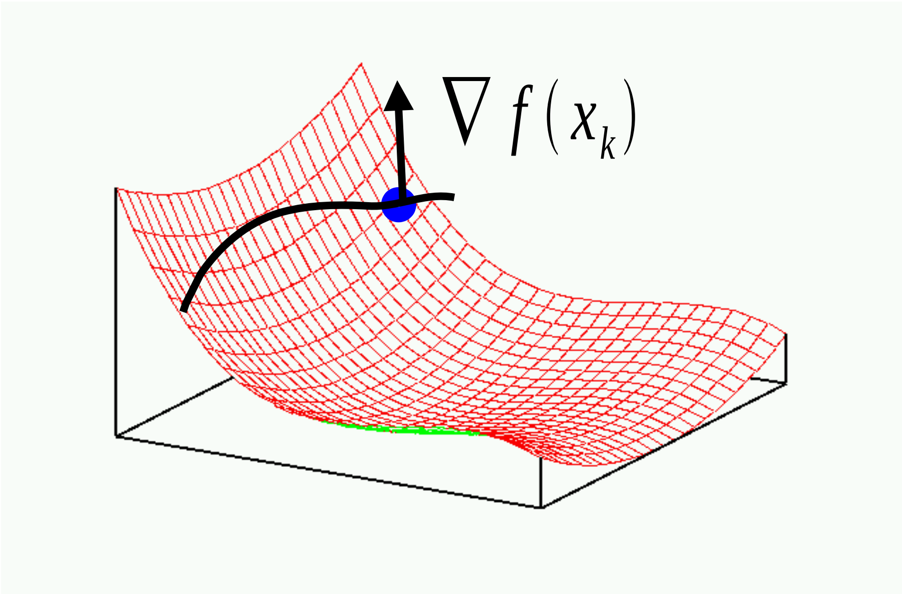
Iterate until we are satisfied.

Step 1: Choose search direction

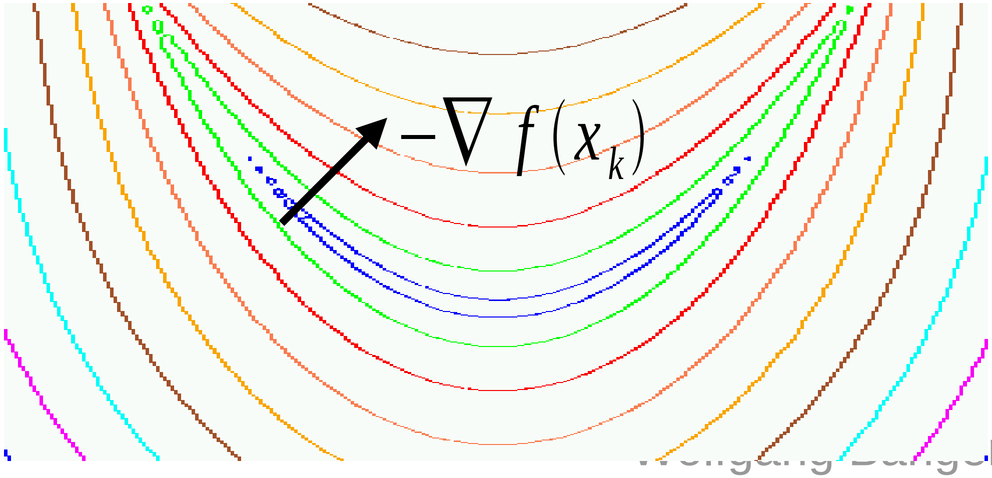
Conditions for a useful search direction:

Minimization function should be decreased in this direction:

$$p_k \cdot \nabla f(x_k) \leq 0$$



Search direction should lead to the minimum as straight as possible



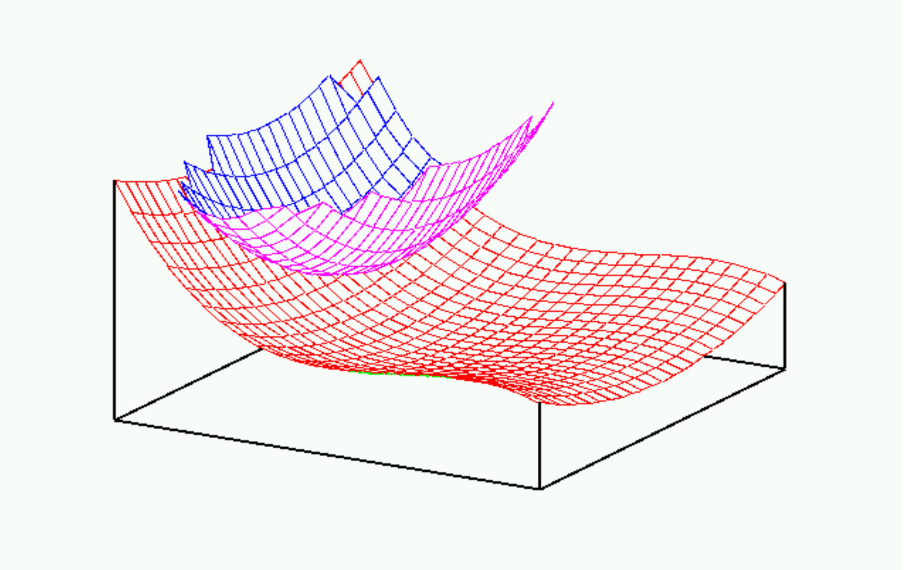
Step 1: Choose search direction

Basic assumption: We can usually only expect to know the minimization function $f(x_k)$ *locally* at x_k .
That means that we can only evaluate

$$f(x_k) \qquad \nabla f(x_k) = g_k \qquad \nabla^2 f(x_k) = H_k \qquad \ldots$$

For a search direction, try to model f in the vicinity of x_k by a Taylor series:

$$\begin{aligned} f(x_k + p_k) \approx\ & f(x_k) \\ & + g_k^T p_k \\ & + \frac{1}{2} p_k^T H_k p_k + \ldots \end{aligned}$$



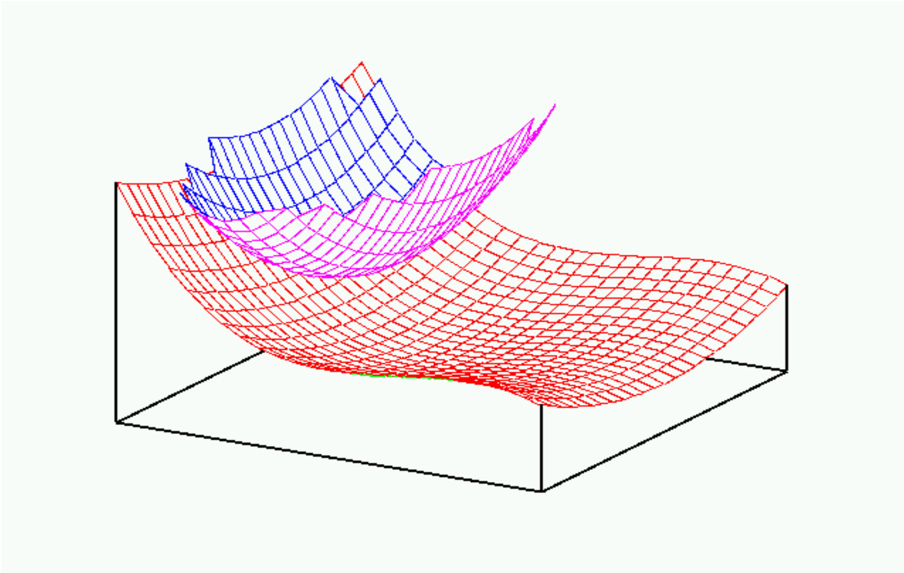
Step 1: Choose search direction

Goal: Approximate $f(\cdot)$ in the vicinity of x_k by a model

$$f(x_k+p) \approx m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T H_k p + \ldots$$

with $f(x_k)=f_k \qquad \nabla f(x_k)=g_k \qquad \nabla^2 f(x_k)=H_k \qquad \ldots$

Then: Choose that direction p_k that minimizes the model $m_k(p)$



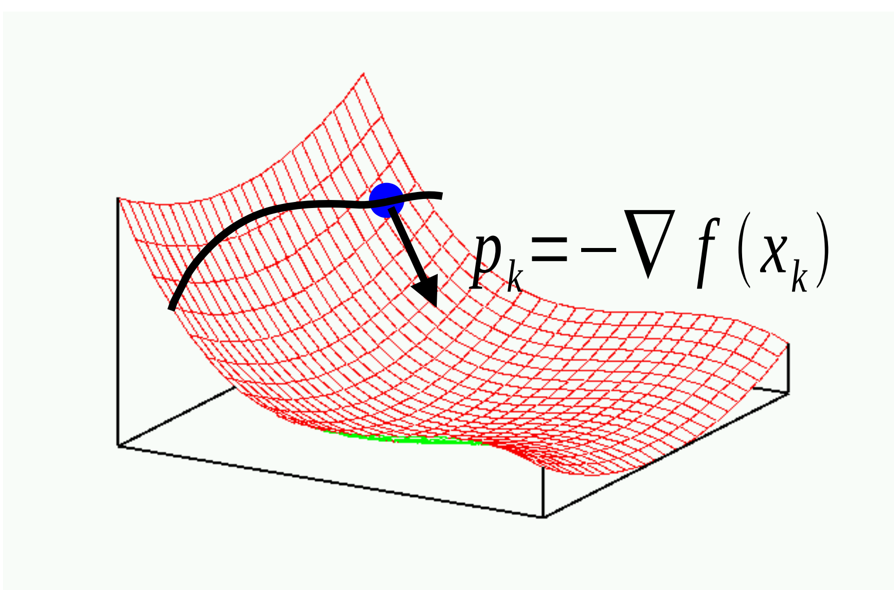
Step 1: Choose search direction

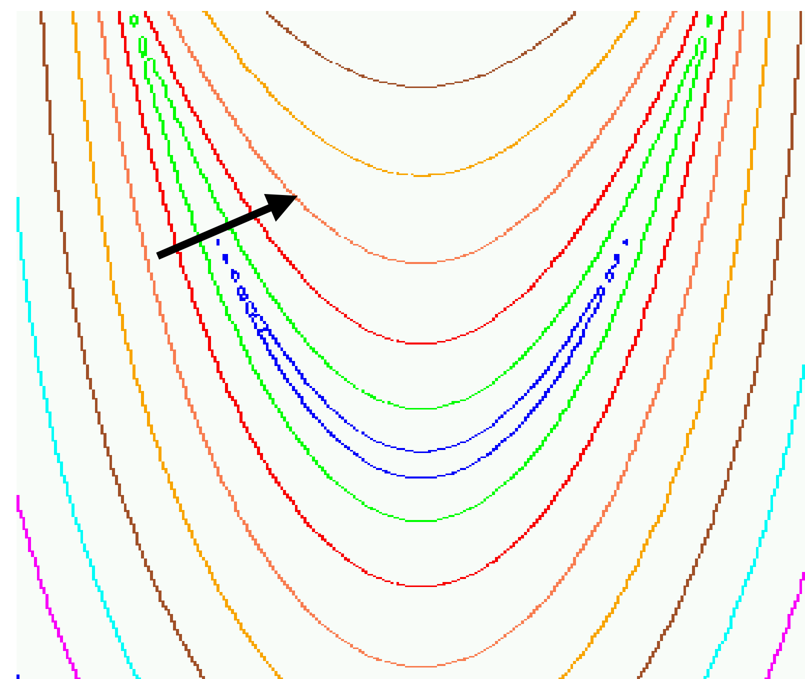
Method 1 (*Gradient method, Method of Steepest Descent*):

search direction is minimizing direction of *linear model*

$$f(x_k+p) \approx f_k + g_k^T p = m_k(p)$$

$$p_k = -g_k$$





Step 1: Choose search direction

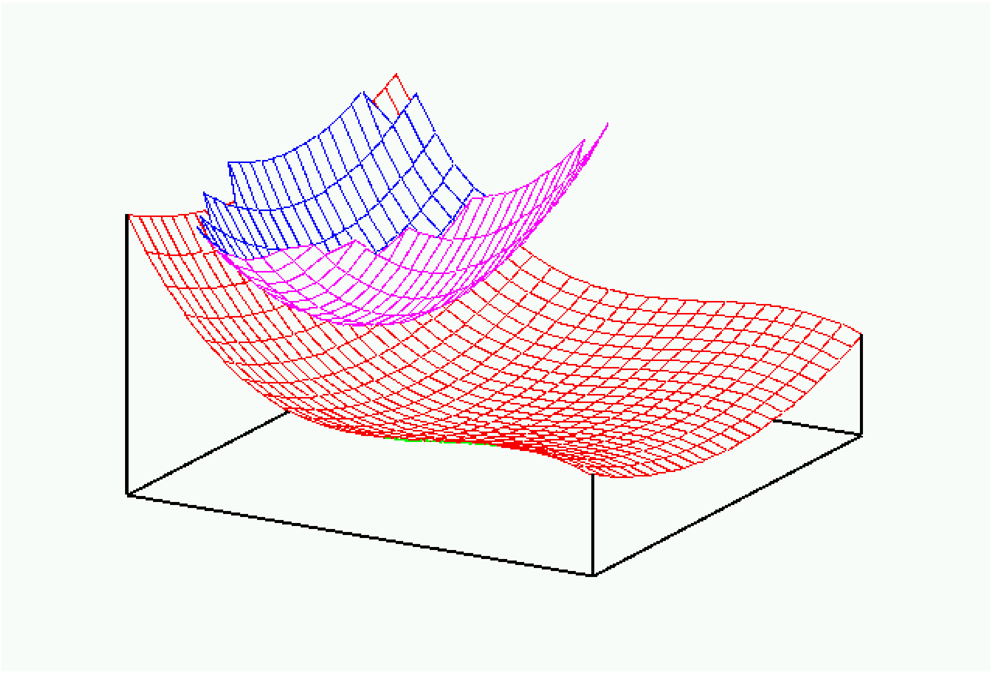
Method 2 (*Newton's method*):

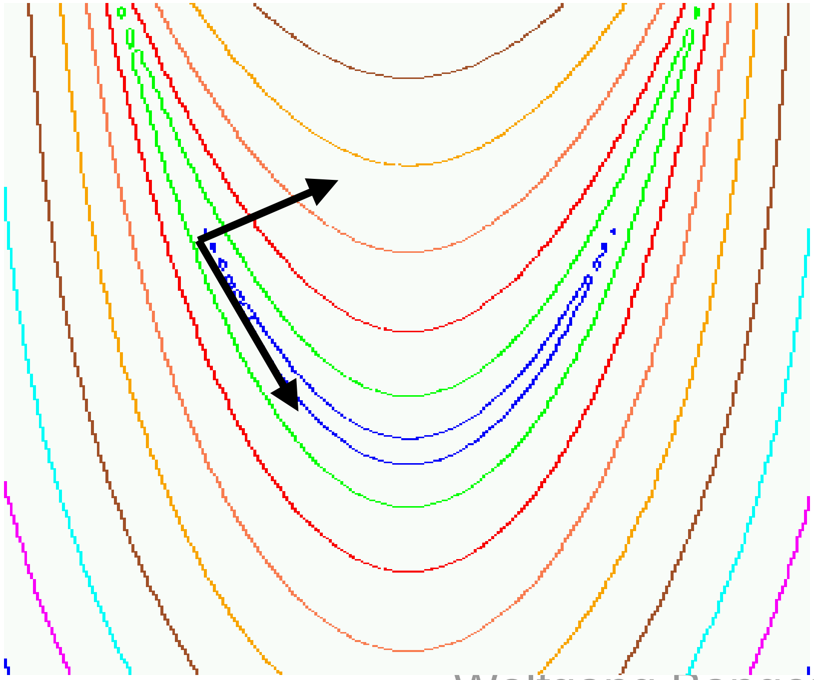
search direction is to the minimum of the *quadratic model*

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T H_k p$$

Minimum is characterized by

$$\frac{\partial m_k(p)}{\partial p} = g_k + H_k p = 0 \quad \rightarrow \quad p_k = -H_k^{-1} g_k$$





Step 1: Choose search direction

Method 2 (*Newton's method*) -- alternative viewpoint:

Newton step is also generated when applying Newton's method for the root-finding problem (*F(x)=0*) to the necessary optimality condition:

$$\nabla f(x^*) = 0$$

Linearize necessary condition around x_k:

$$0 = \nabla f(x^*) = \underbrace{\nabla f(x_k)}_{g_k} + \underbrace{\nabla^2 f(x_k)}_{H_k} \underbrace{(x^* - x_k)}_{p_k} + \ldots$$

$$p_k = -H_k^{-1} g_k$$

Step 1: Choose search direction

Method 3 (*A third order method*):

The search direction is to the minimum of the *cubic model*

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T H_k p + \frac{1}{6}\left[\frac{\partial^3 f}{\partial x_l \partial x_m \partial x_n}\right]_k p_l p_m p_n$$

Minimum is characterized by the quadratic equation

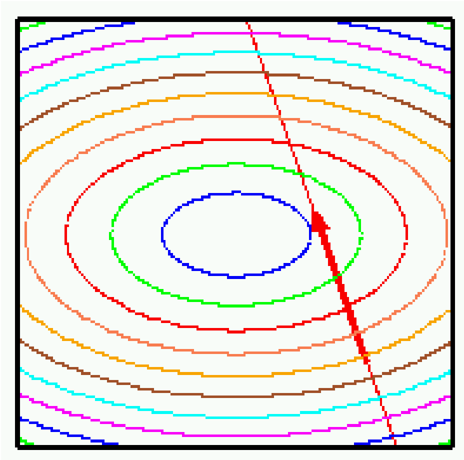
$$\frac{\partial m_k(p)}{\partial p} = g_k + H_k p + \frac{1}{2}\left[\frac{\partial^3 f}{\partial x_l \partial x_m \partial x_n}\right]_k p_l p_m = 0 \quad \rightarrow \quad p_k = ???$$

But: There is no practical way to compute the solution of this equation for problems with more than one variable.

Step 2: Determination of Step Length

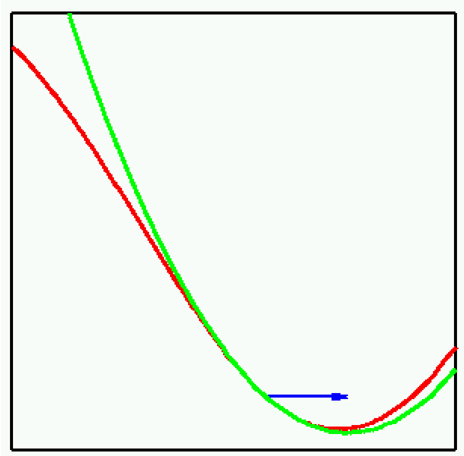
Once the search direction is known, compute the update by choosing a step length α_k and set

$$x_{k+1} = x_k + \alpha_k p_k$$



Determine the step length by solving the 1-d minimization problem (*line search*):

$$\alpha_k = \arg\min_{\alpha} f(x_k + \alpha p_k)$$



For Newton's method: If the quadratic model is good, then step is good, then take *full step* with $\alpha_k = 1$

Convergence: Gradient method

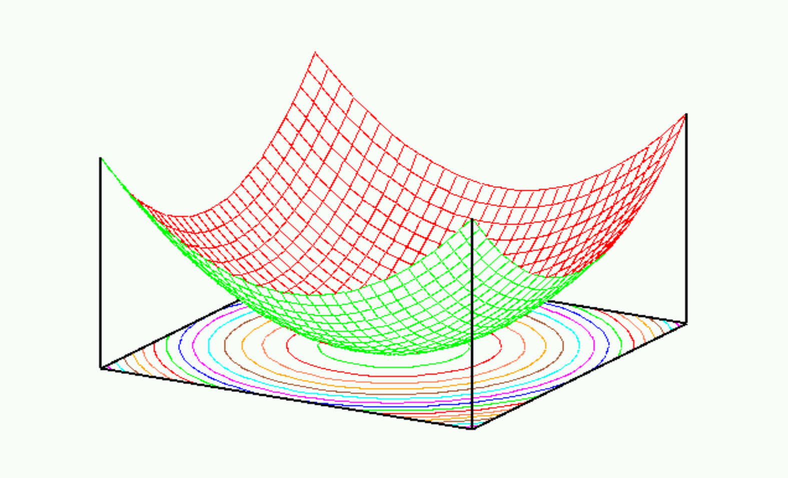
Gradient method converges *linearly*, i.e.

$$\|x_k - x^*\| \;\leq\; C\,\|x_{k-1} - x^*\|$$

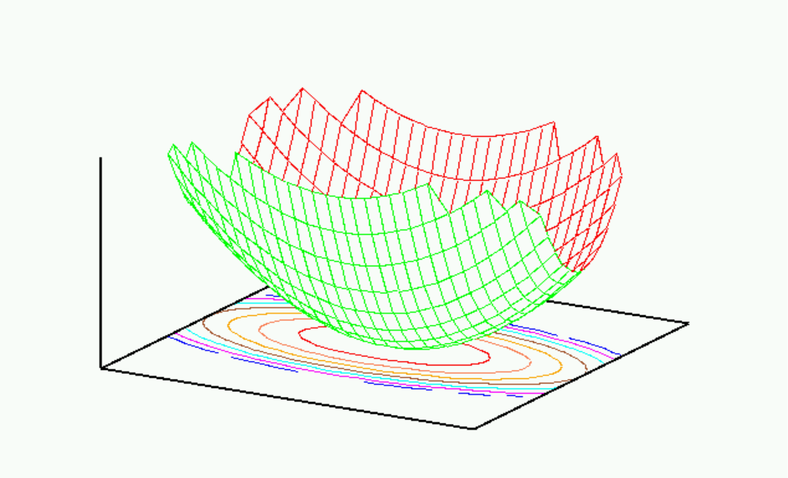
Gain is a fixed factor $C<1$
Convergence can be *very* slow if C close to 1.

Example: If $f(x)=x^THx$, with H positive definite and for optimal line search, then

$$C \approx \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \qquad \{\lambda_i\} = \text{spectrum } H$$



$x^2+y^2 \;\rightarrow\; C=0$



$x^2+5y^2 \;\rightarrow\; C\approx 0.6$

Convergence: Newton's method

Newton's method converges *quadratically*, i.e.

$$\|x_k - x^*\| \leq C \|x_{k-1} - x^*\|^2$$

Optimal convergence order only if step length is 1, otherwise slower convergence (step length is 1 if quadratic model valid!)

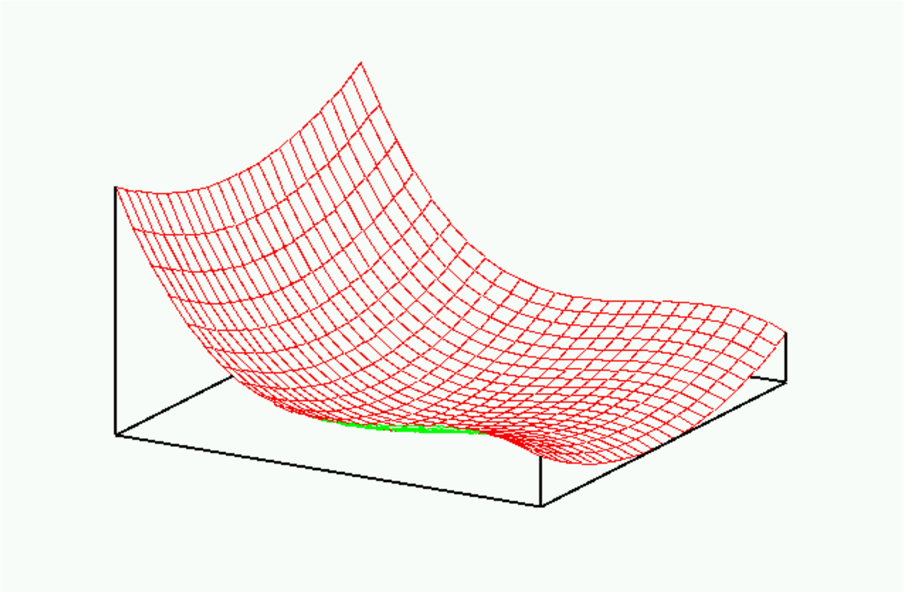
If quadratic convergence: accelerating progress as iterations proceed.

Size of C:

$$C \sim \sup_{x,y} \frac{\left\| \nabla^2 f(x^*)^{-1} \left(\nabla^2 f(x) - \nabla^2 f(y) \right) \right\|}{\|x-y\|}$$

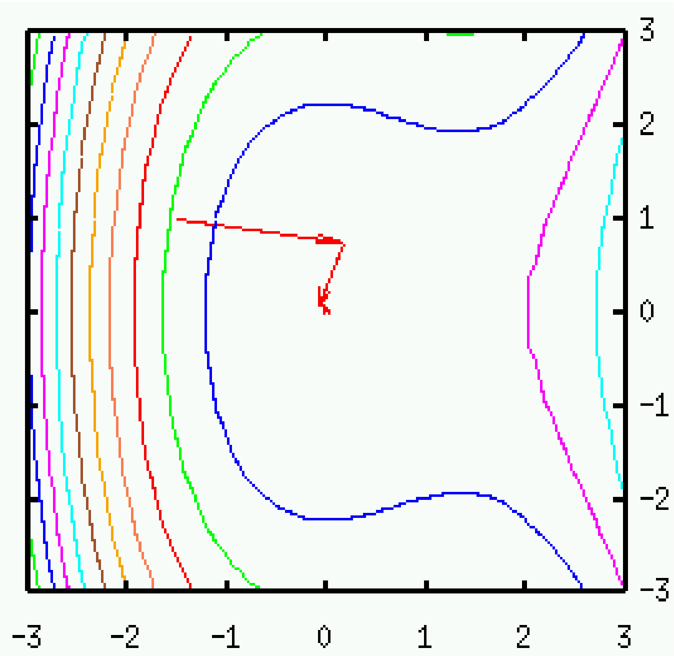
C measures size of nonlinearity beyond quadratic part.

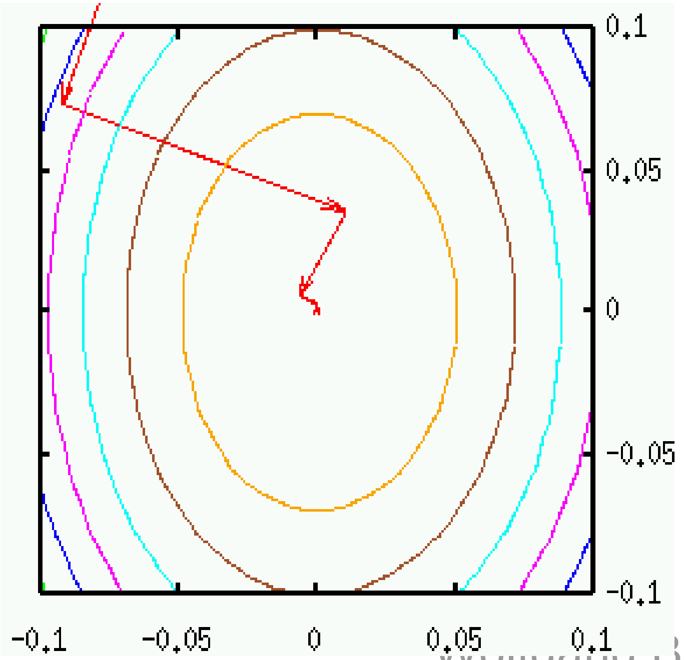
Example 1: Gradient method



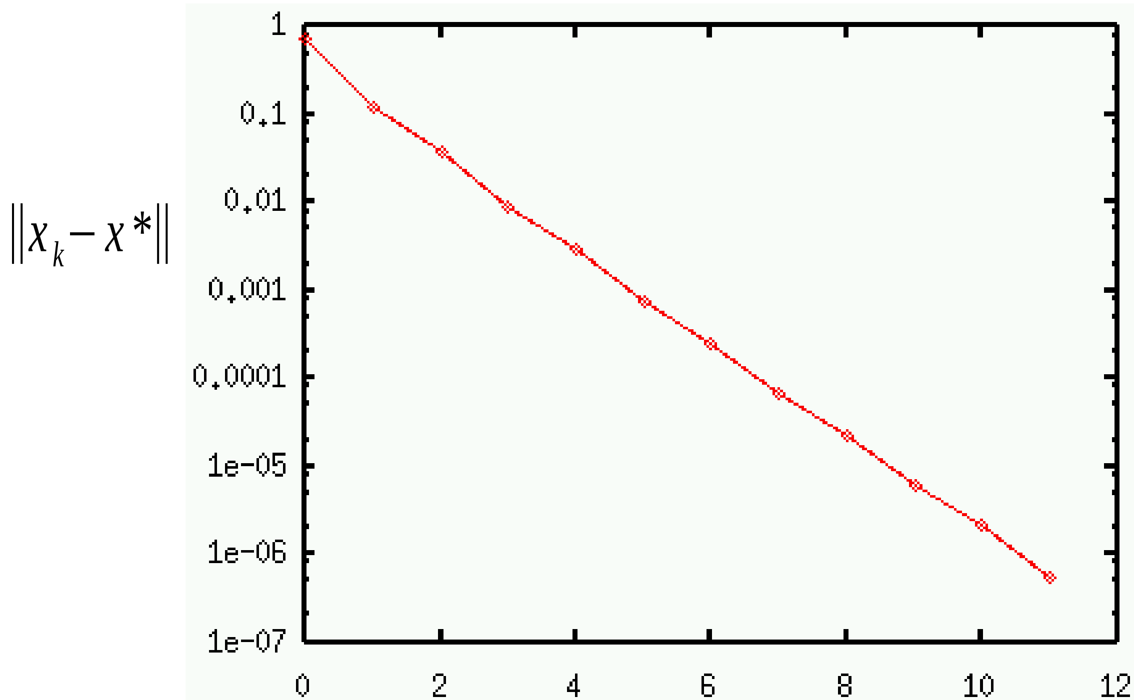
$$f(x,y) = -x^3 + 2x^2 + y^2$$

Local minimum at x=y=0,
saddle point at x=4/3, y=0





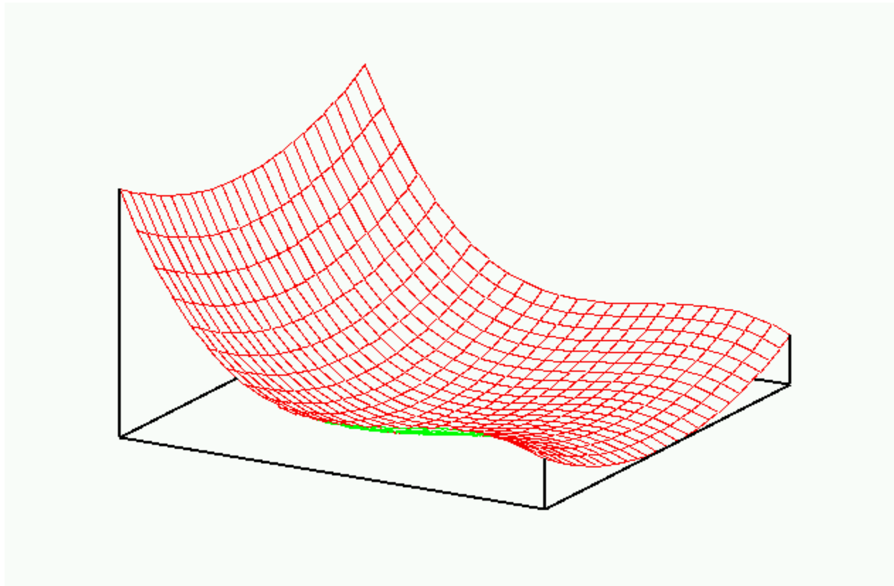
Example 1: Gradient method



Convergence of gradient method:
Converges quite fast, with *linear* rate
Mean value of convergence constant C : 0.28
At (x=0,y=0), there holds

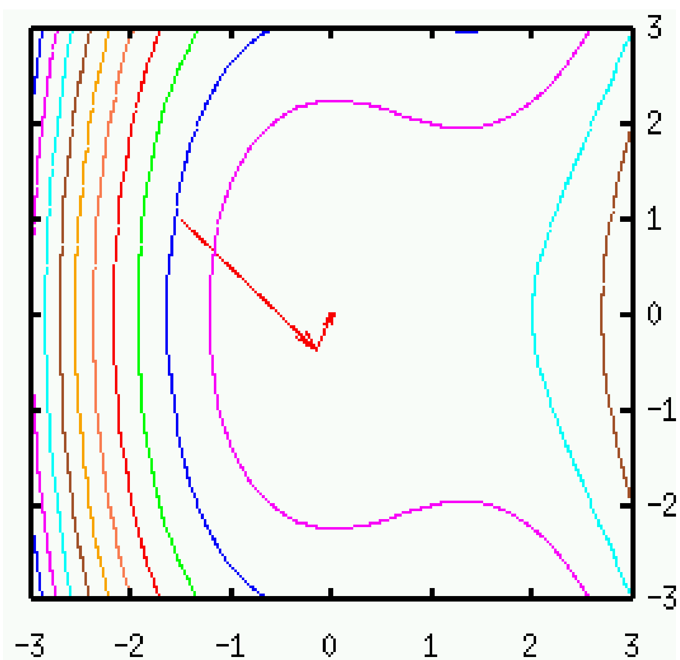
$$\nabla^2 f(0,0) \sim \{\lambda_1 = 4, \lambda_2 = 2\} \qquad C \approx \frac{4-2}{4+2} \approx 0.33$$

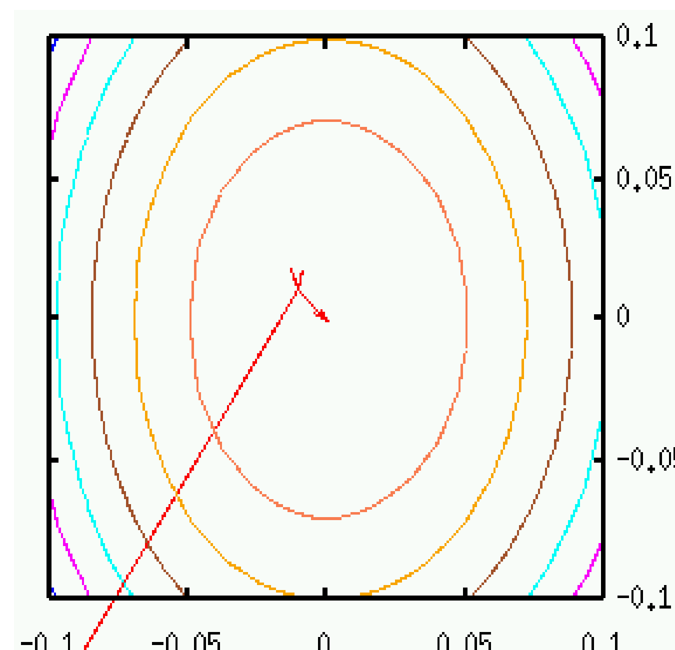
Example 1: Newton's method



$$f(x,y) = -x^3 + 2x^2 + y^2$$

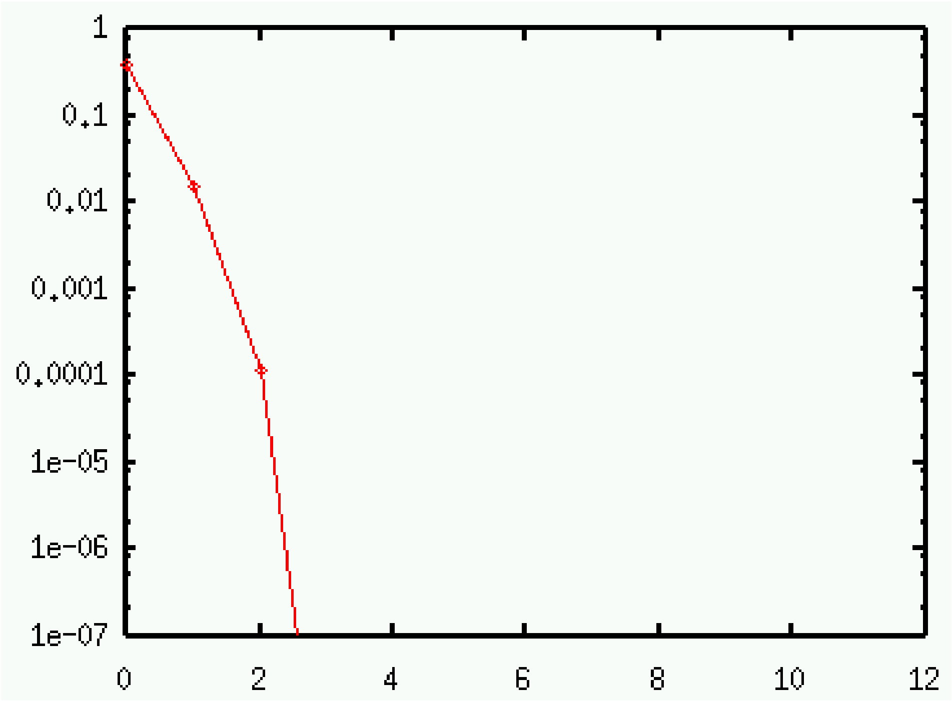
Local minimum at x=y=0, saddle point at x=4/3, y=0





Example 1: Newton's method

$\|x_k - x^*\|$

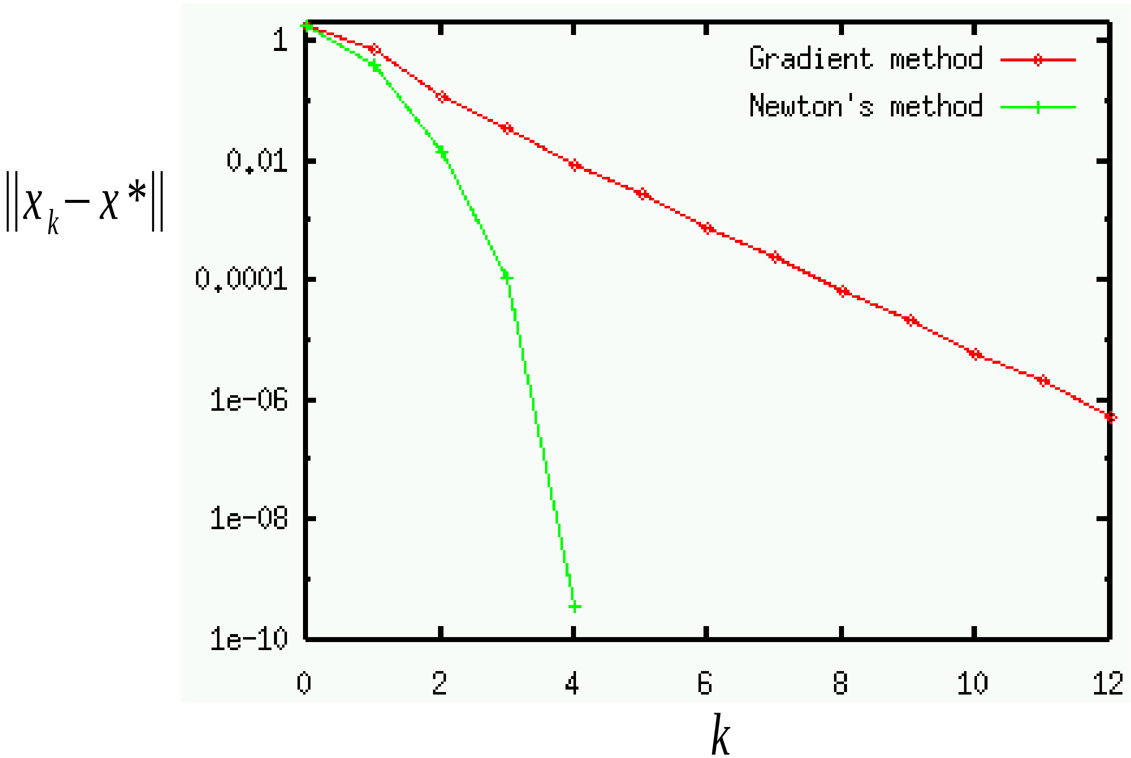


Convergence of Newton's method:
Converges very fast, with *quadratic* rate
Mean value of convergence constant C : 0.15

$$\|x_k - x^*\| \leq C \|x_{k-1} - x^*\|^2$$

Theoretical estimate yields C=0.5

Example 1: Comparison between methods

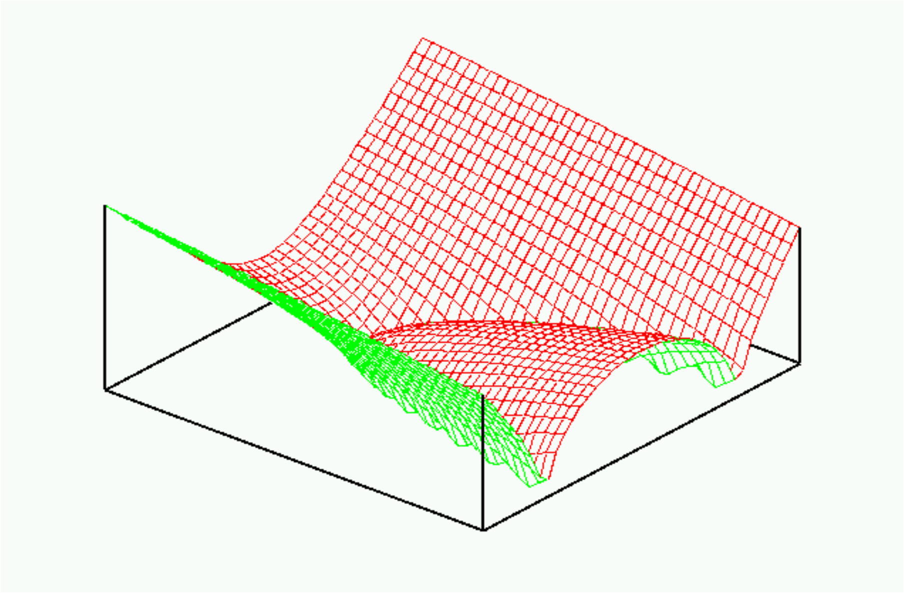


Newton's method much faster than gradient method

Newton's method superior for high accuracy due to higher order of convergence

Gradient method simple but converges in a reasonable number of iterations as well

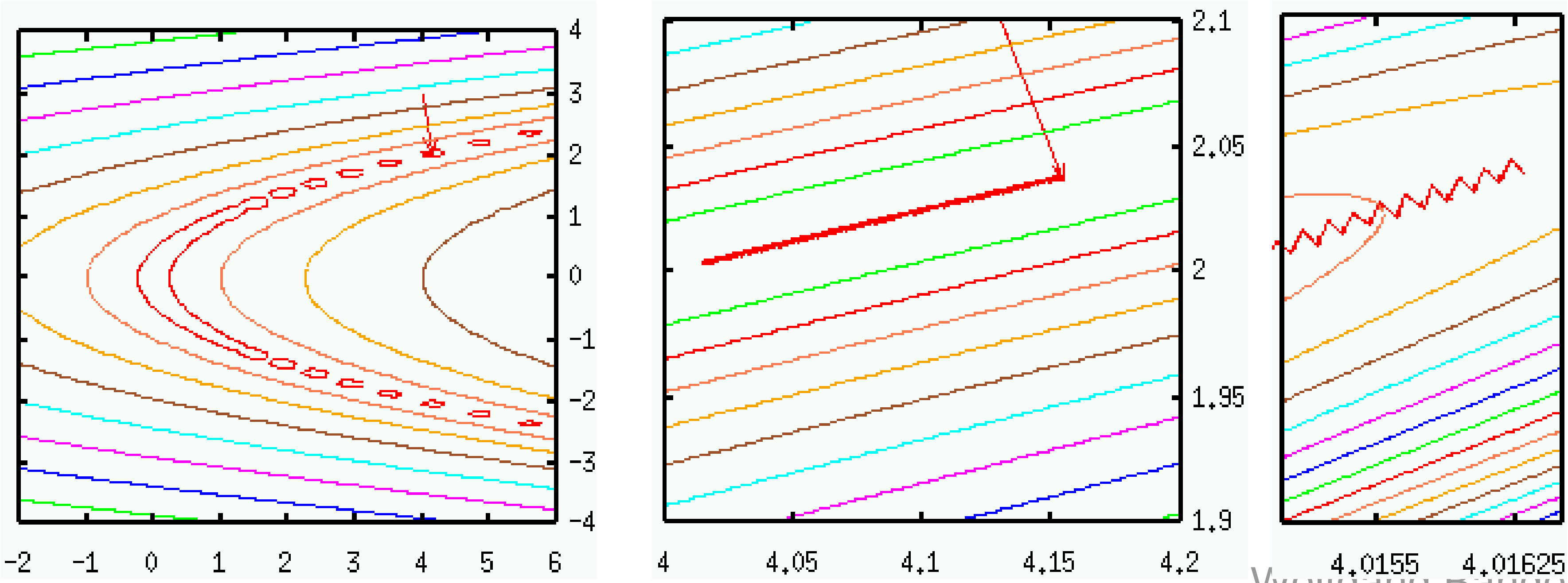
Example 2: Gradient method



$$f(x,y)=\sqrt[4]{\left((x-y^2)^2+\frac{1}{100}\right)+\frac{1}{100}y^2}$$

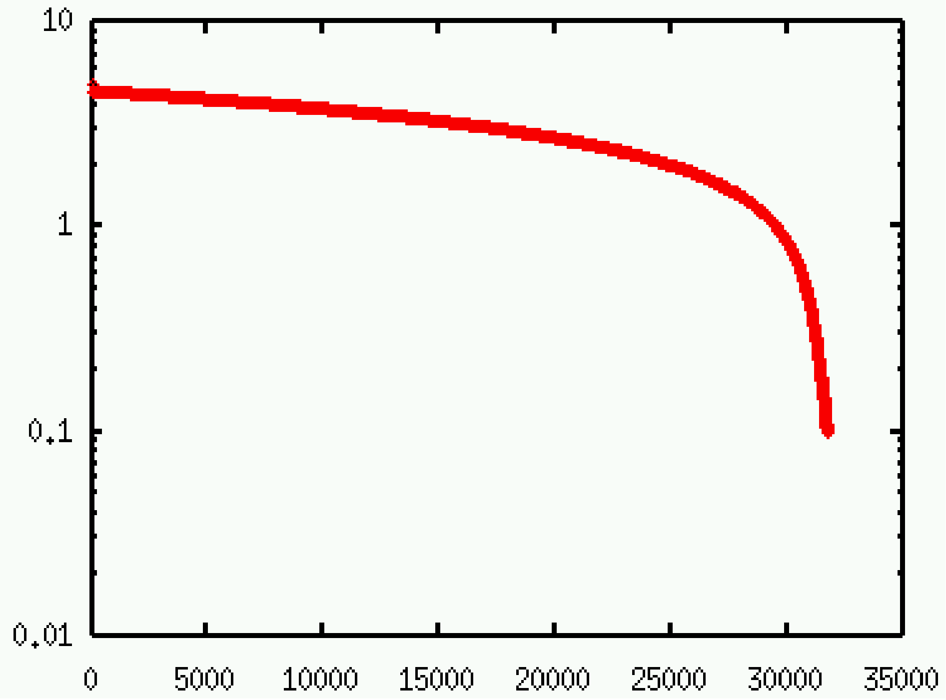
(*Banana valley* function)

Global minimum at x=y=0



Example 2: Gradient method

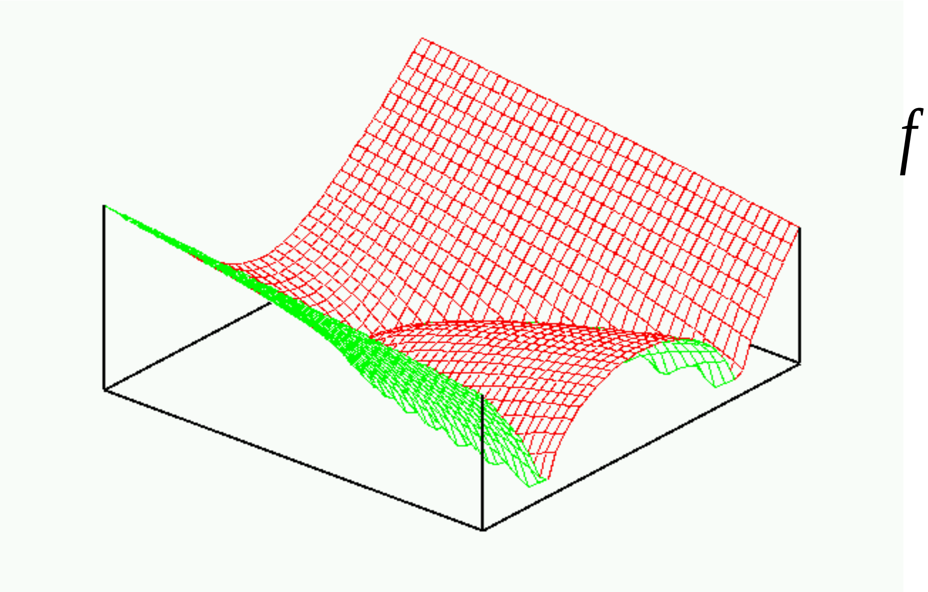
$\|x_k - x^*\|$



Convergence of gradient method:
Needs almost 35,000 iterations to come closer than 0.1 to the solution!
Mean value of convergence constant C : 0.99995
At (x=4,y=2), there holds

$$\nabla^2 f(4,2) \sim \{\lambda_1 = 0.1, \lambda_2 = 268\} \qquad C \approx \frac{268-0.1}{268+0.01} \approx 0.9993$$

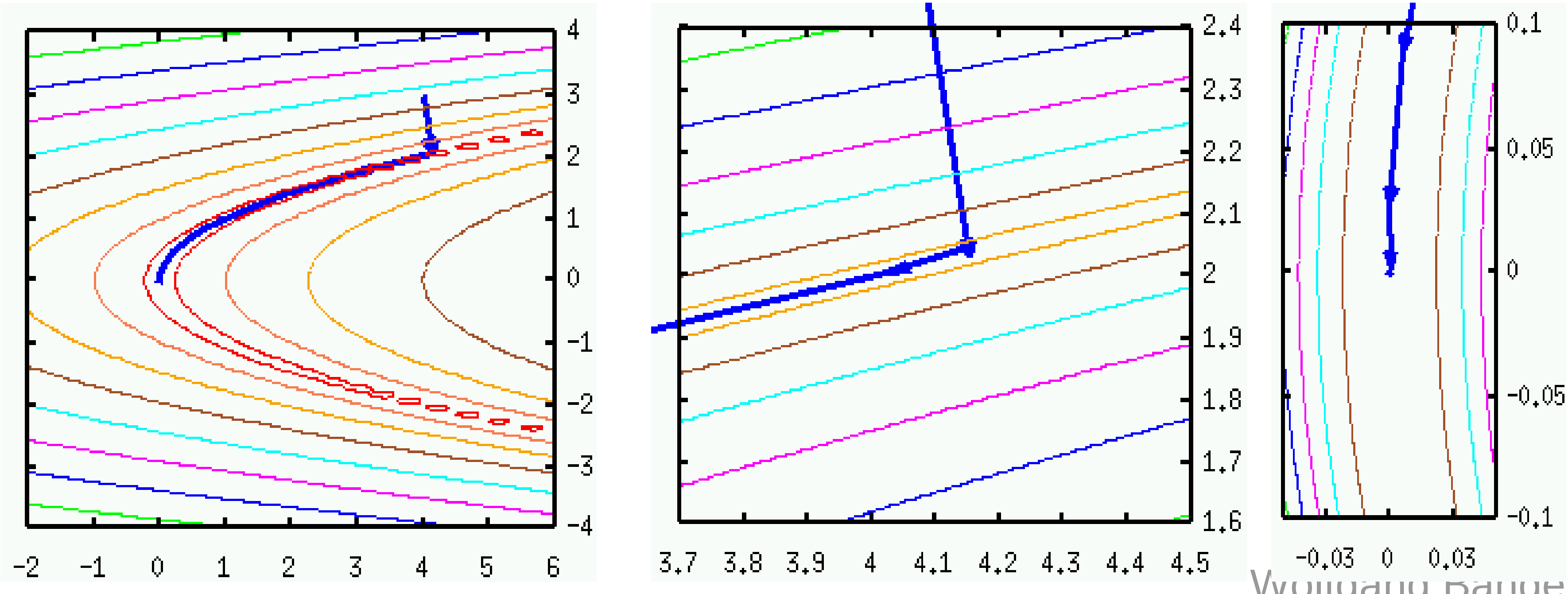
Example 2: Newton's method



$$f(x,y) = \sqrt[4]{\left((x-y^2)^2 + \frac{1}{100}\right) + \frac{1}{100}y^2}$$

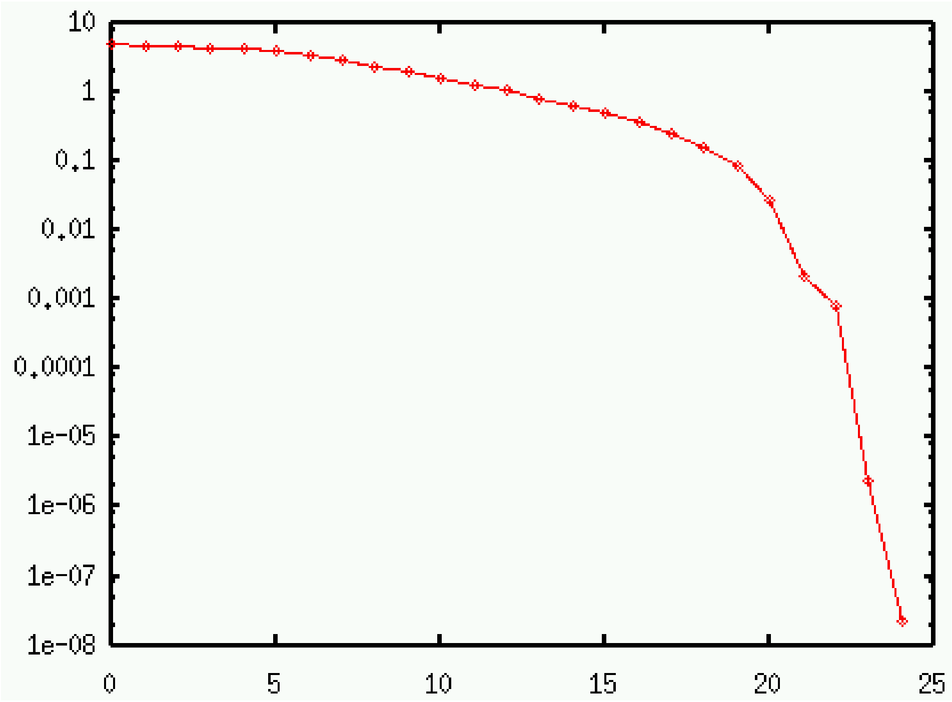
(*Banana valley* function)

Global minimum at x=y=0



Example 2: Newton's method

$\|x_k - x^*\|$

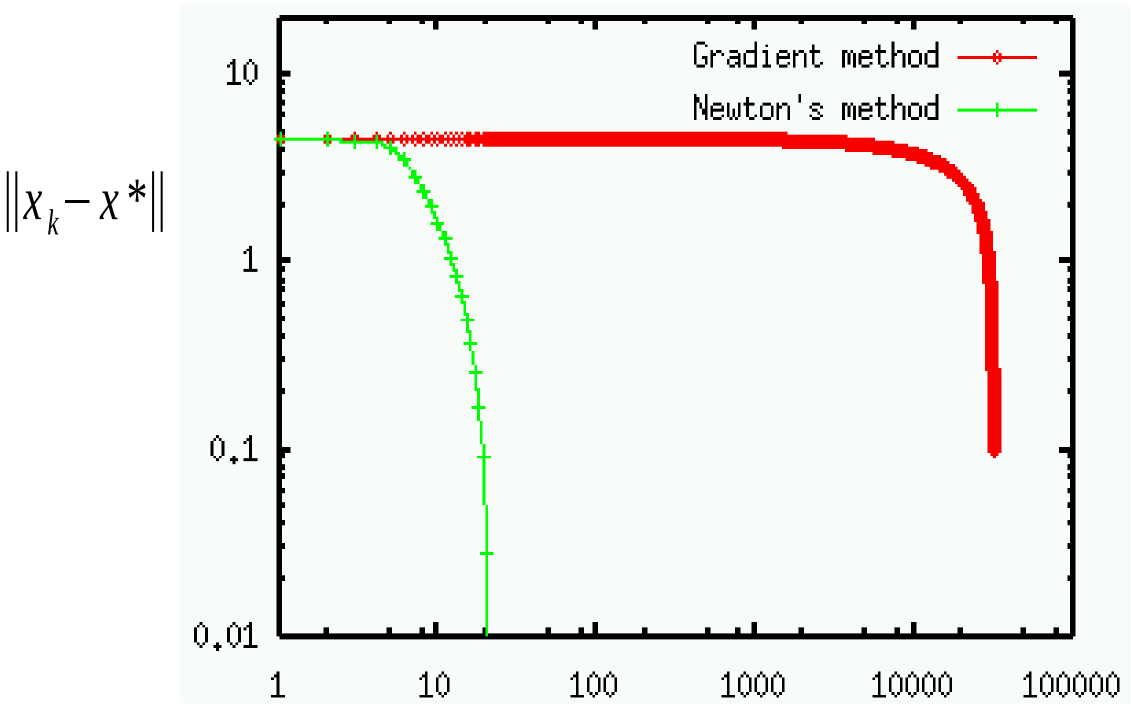


Convergence of Newton's method:
Less than 25 iterations for an accuracy of better than 10^{-7}!

Convergence roughly *linear* for first 15-20 iterations since step length $\alpha_k \neq 1$

Convergence roughly quadratic for last iterations with step length $\alpha_k \approx 1$

Example 2: Comparison between methods



Newton's method **much** faster than gradient method

Newton's method superior for high accuracy (i.e. in the vicinity of the solution) due to higher order of convergence

Gradient method converges too slowly for practical use

Practical line search strategies

Ideally: Use an exact step length determination (*line search*) based on

$$\alpha_k = \arg\min_\alpha f(x_k + \alpha p_k)$$

This is a 1d minimization problem for α, solvable via Newton's method/bisection search/etc.

However: Expensive, may require many function/gradient evaluations.

Instead: Find practical criteria that guarantee convergence but need less function evaluations!

Practical line search strategies

Strategy: Find practical criteria that guarantee convergence but need less evaluations.

Rationale:

- Near the optimum, quadratic approximation of *f* is valid
  → take full steps (step length 1) there
- Line search only necessary far away from the solution
- If close to solution, need to try α=1 first

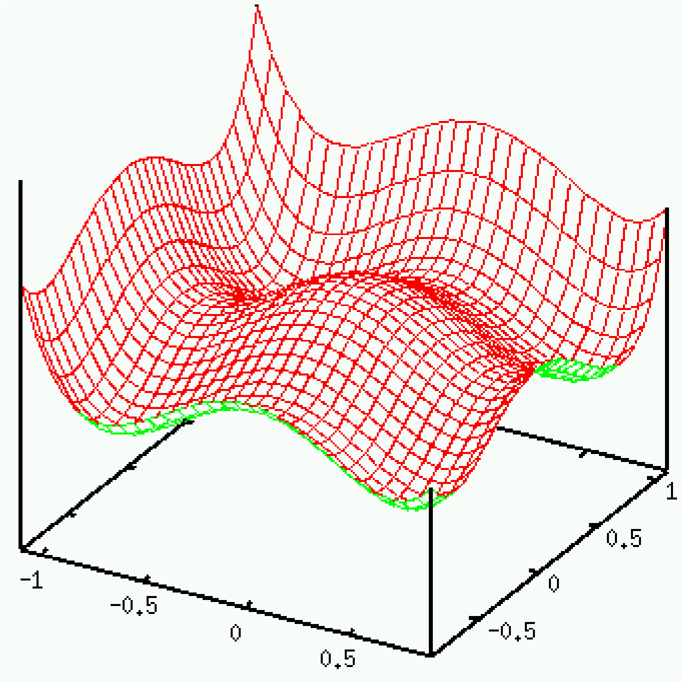
Consequence:

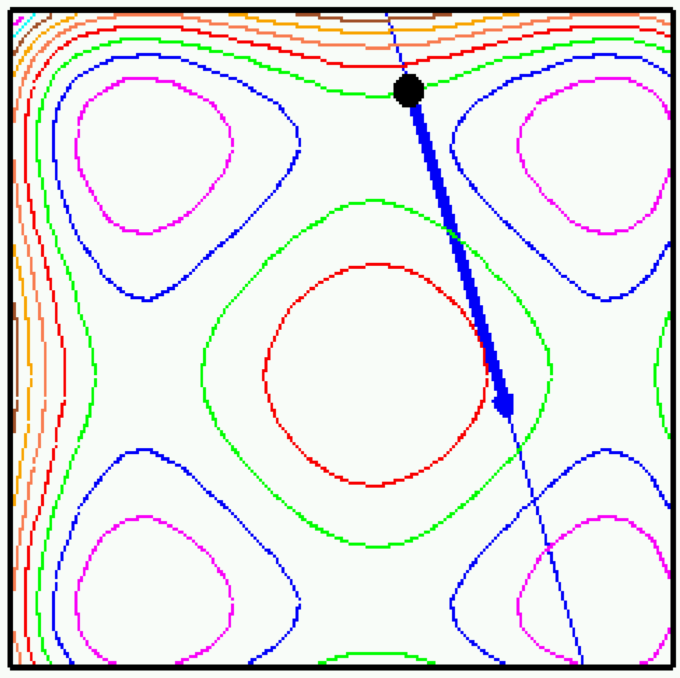
- Near solution, quadratic convergence of Newton's method is retained
- Far away, convergence is slower in any case.

Practical line search strategies

Practical strategy: Use an inexact line search that:

- finds a reasonable approximation to the exact step length
- chosen step length guarantees a *sufficient decrease* in *f(x)*;
- chooses full step length 1 for Newton's method whenever possible.



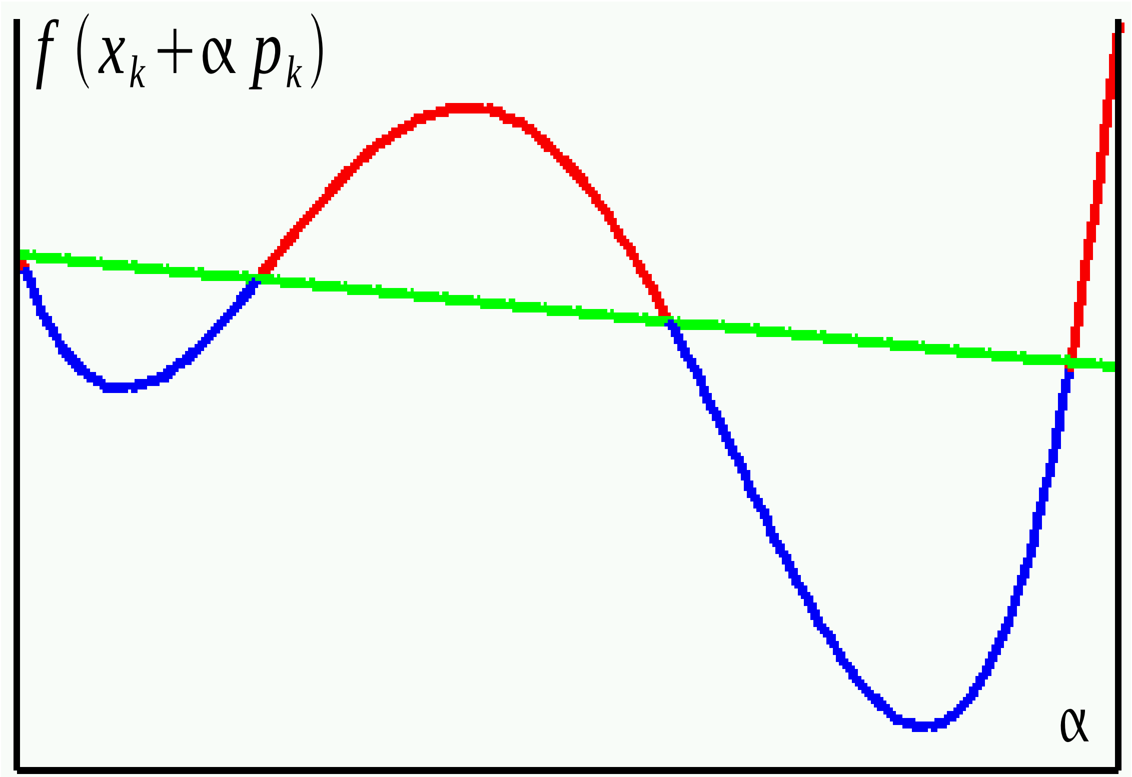


$$f(x, y) = x^4 - x^2 + y^4 - y^2$$

Practical line search strategies

Wolfe condition 1 ("sufficient decrease" condition):
Require step lengths to produce a sufficient decrease

$$
\begin{aligned}
f(x_k+\alpha p_k) &\le f(x_k) + c_1 \alpha \left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=0} \\
&= f_k + c_1 \alpha \nabla f_k \cdot p_k
\end{aligned}
$$



Necessary:

$$0<c_1<1$$

Typical values:

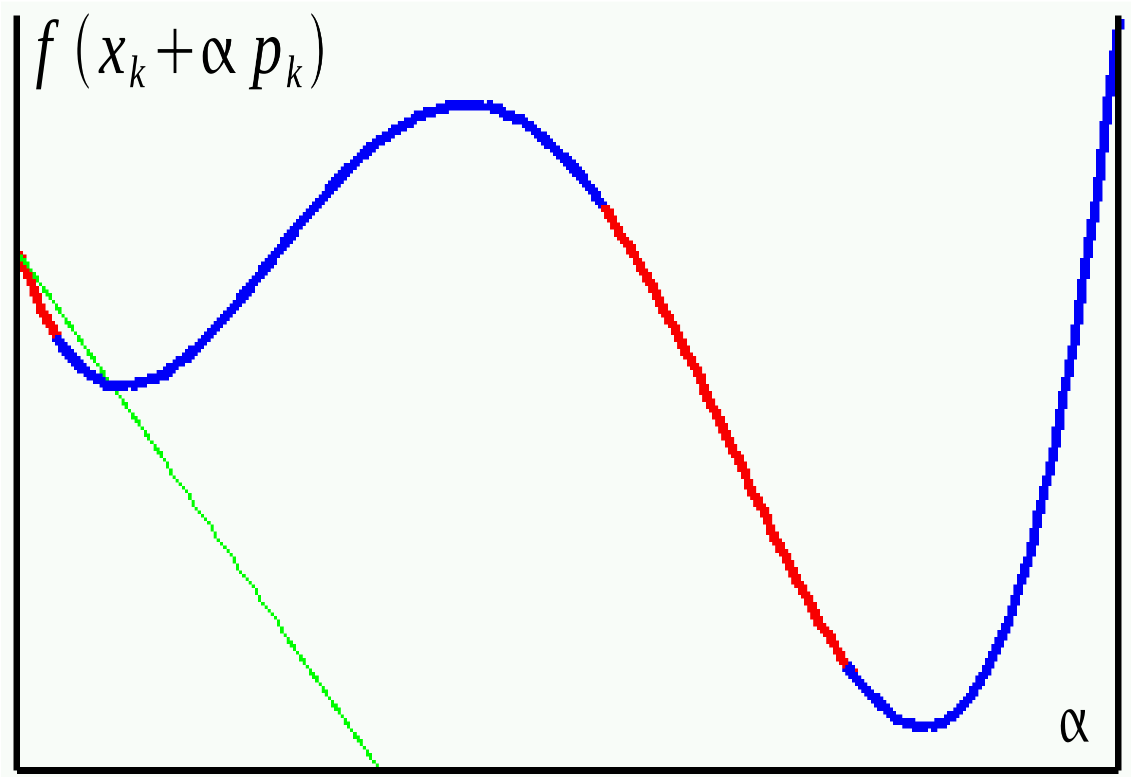
$$c_1=10^{-4}$$

i.e.: only very small decrease mandated

Practical line search strategies

Wolfe condition 2 ("curvature" condition):
Require step lengths where *f* has shown sufficient curvature upwards

$$\nabla f(x_k+\alpha p_k)\cdot p_k = \left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=\alpha_k} \geq c_2\left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=0} = c_2 \nabla f_k \cdot p_k$$



Necessary:

$$0<c_1<c_2<1$$

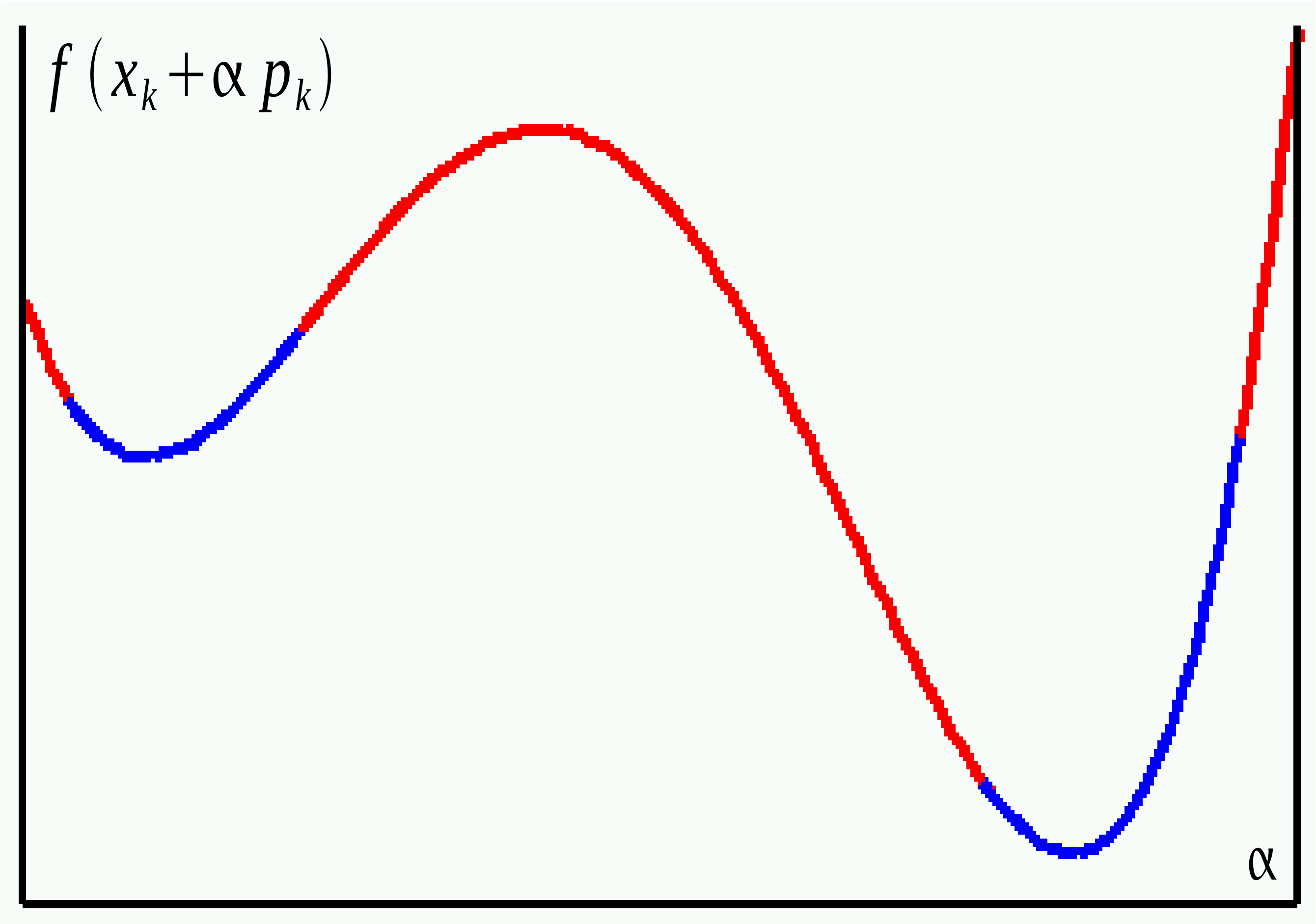
Typical:

$$c_2=0.9$$

Rationale: Exclude too small step lengths

Wolfe conditions

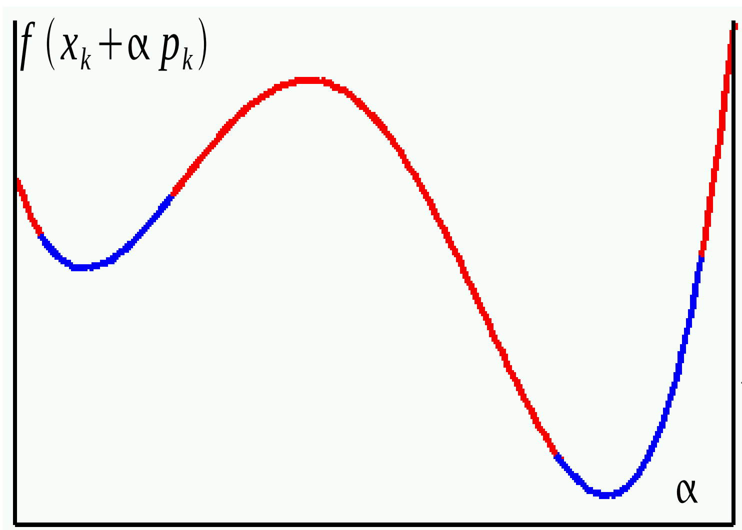
Conditions 1 and 2 usually yield reasonable ranges for the step lengths, but do not guarantee optimal ones

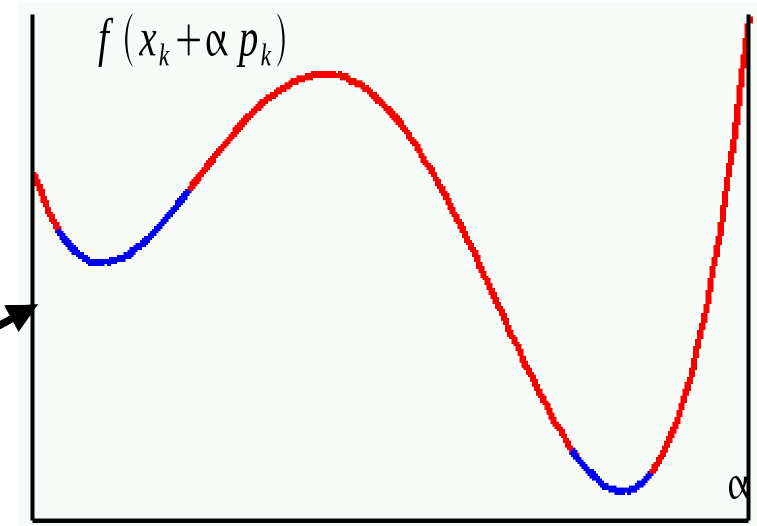


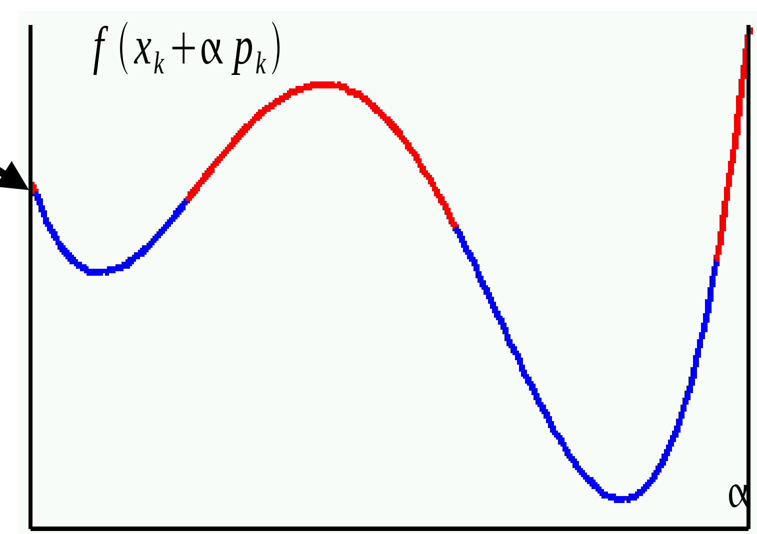
Practical line search strategies - Alternatives

Strict Wolfe conditions:

$$\left|\left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=\alpha_k}\right| \le c_2 \left|\left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=0}\right|$$







Goldstein conditions:

$$f(x_k+\alpha p_k) \ge f(x_k) + (1-c_1)\alpha\left[\frac{\partial f(x_k+\alpha p_k)}{\partial \alpha}\right]_{\alpha=0}$$

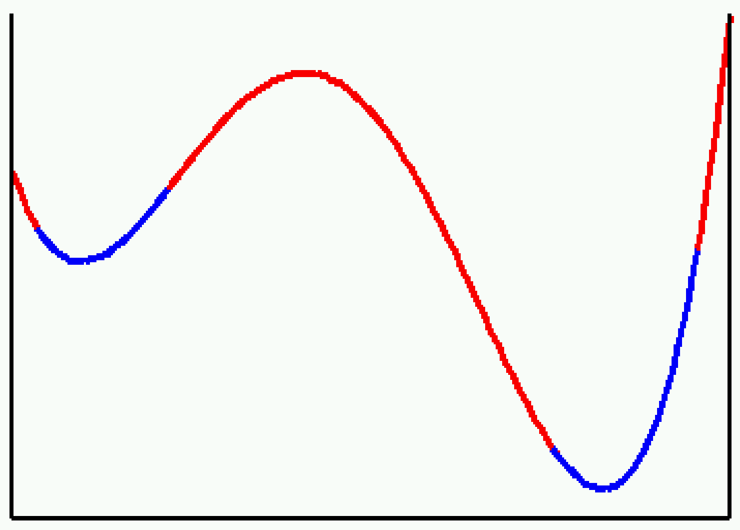
Practical line search strategies

Conditions like the ones above tell us whether a given step length is acceptable or not.

In practice, don't try too many step lengths – checking the conditions involves function evaluations of *f(x).*

Typical strategy ("Backtracking line search"):

1. Start with a trial step length $\alpha_t = \bar{\alpha}$
   (for Newton's method: $\bar{\alpha} = 1$)
2. Verify acceptance conditions for this α_t
3. If yes: $\alpha_k = \alpha_t$
4. If no: $\alpha_t = c\,\alpha_t,\ c < 1$ and go to 2.



Note: A typical reduction factor is $c = \frac{1}{2}$

Practical line search strategies

An alternative strategy ("Interpolating line search"):

- Start with $\alpha_t^{(0)}=\bar{\alpha}=1$, set $i=0$
- Verify acceptance conditions for $\alpha_t^{(i)}$
- If yes: $\alpha_k=\alpha_t^{(i)}$
- If no:
  - let $\phi_k(\alpha)=f(x_k+\alpha p_k)$
  - from evaluating the sufficient decrease condition

$$f(x_k+\alpha_t^{(i)} p_k) \quad \leq \quad f_k \; + \; c_1 \alpha_t^{(i)} \nabla f_k \cdot p_k$$

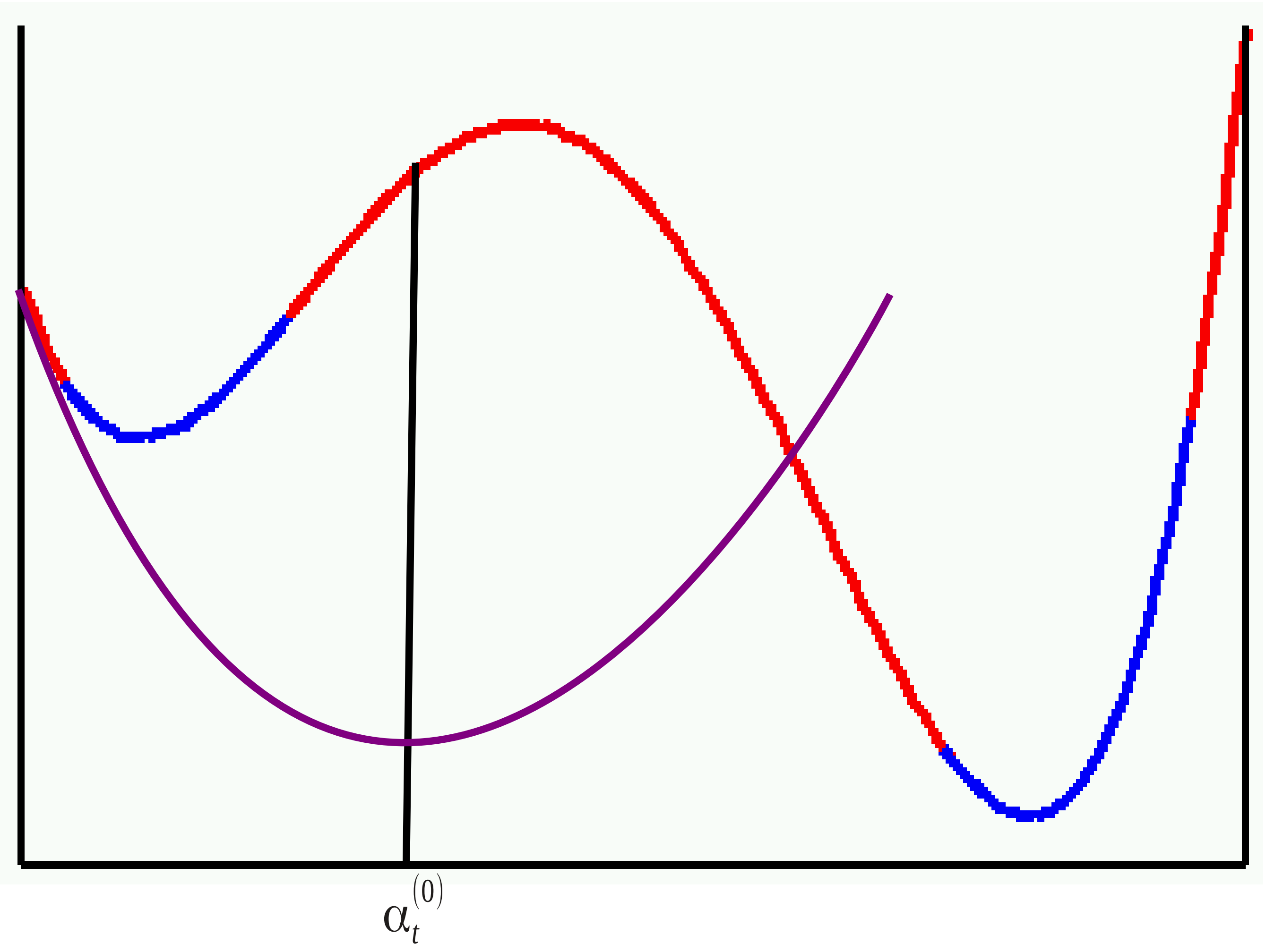
  we already know $\phi_k(0)=f(x_k)$, $\phi_k'(0)=\nabla f_k \cdot p_k=g_k \cdot p_k$
  and $\phi_k(\alpha_t^{(i)})=f(x_k+\alpha_t^{(i)} p_k)$

  - if $i=0$ then choose $\alpha_t^{(i+1)}$ as minimizer of the quadratic function that interpolates $\phi_k(0), \phi'_k(0), \phi_k(\alpha_t^{(i)})$
  - if $i>0$ then choose $\alpha_t^{(i+1)}$ as the minimizer of the cubic function that interpolates $\phi_k(0), \phi'_k(0), \phi_k(\alpha_t^{(i)}), \phi_k(\alpha_t^{(i-1)})$

Practical line search strategies

An alternative strategy ("Interpolating line search"):

Step 1: Quadratic interpolation



Practical line search strategies

An alternative strategy ("Interpolating line search"):

Step 2 and following: Cubic interpolation