NTNU | Norwegian University of Science and Technology

TTK4135 – Lecture 14
Line search

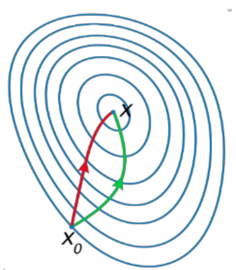
Lecturer: Lars Imsland

Learning goal Ch. 2, 3 and 6: Understand this slide

Line-search unconstrained optimization

$$\min_x f(x)$$

1. Initial guess x_0
2. While termination criteria not fulfilled
   a) Find descent direction p_k from x_k
   b) Find appropriate step length α_k ; set $x_{k+1} = x_k + \alpha_k p_k$
   c) $k = k+1$
3. $x_M = x^*$? (possibly check sufficient conditions for optimality)



A comparison of steepest descent and Newton's method. Newton's method uses curvature information to take a more direct route. (wikipedia.org)

Termination criteria:

Stop when first of these become true:

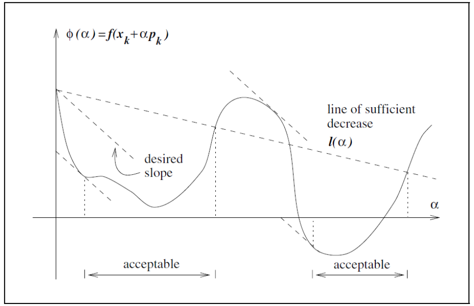
- $\|\nabla f(x_k)\| \leq \epsilon$ (necessary condition)
- $\|x_k - x_{k-1}\| \leq \epsilon$ (no progress)
- $\|f(x_k) - f(x_{k-1})\| \leq \epsilon$ (no progress)
- $k \leq k_{\max}$ (kept on too long)

Descent directions:

- Steepest descent
  $p_k = -\nabla f(x_k)$
- Newton
  $p_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$
- Quasi-Newton
  $p_k = -B_k^{-1} \nabla f(x_k)$
  $B_k \approx \nabla^2 f(x_k)$

How to calculate derivatives – Ch. 8

Step length line search (Wolfe):



How many iterations? (Convergence rates)

Outline today: Line search

Objective of line search: make gradient algorithms work when you start far away from optimum

- These algorithms are sometimes called globalization strategies
- Two basic globalization strategies: line search (Ch. 3) and trust-region (Ch. 4, not syllabus)
  - Note again: “globalization” does not imply that we search for global optimum, but we make the algorithm work far from a (local or global) optimum!

Line search elements:

- Conditions on step-length: Wolfe conditions
- Step-length computation

Hessian modification for Newton

Reference: N&W Ch.3-3.1, 3.4, 3.5

Unconstrained opt. iterates: $x_{k+1} = x_k + \underbrace{\alpha_k}_{\text{step length}} \underbrace{p_k}_{\text{descent direction}}$

$$\min_x f(x)$$

Gradient search directions

$$\rightarrow \quad p_k = -B_k^{-1} \underbrace{\nabla f(x_k)}_{\nabla f_k}$$

$B_k = I$: Steepest descent

$\rightarrow$ $B_k = \nabla^2 f_k$: Newton

$B_k \approx \nabla^2 f_k$: Quasi-Newton

B_k is symmetric

Observation: $B_k > 0 \Rightarrow p_k = -B_k^{-1} \nabla f_k$ is a descent direction

Proof: $p_k^T \nabla f_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0$, since $B_k^{-1} > 0$

Quadratic approximation to objective function

$$f(x_k + p) \approx m_k(p) = f(x_k) + p^\top \nabla f(x_k) + \frac{1}{2} p^\top \nabla^2 f(x_k) p$$

Minimize approximation:

$$\nabla_p m_k(p) = 0 \Rightarrow p_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

"Newton step":

$$x_{k+1} = x_k + p_k = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

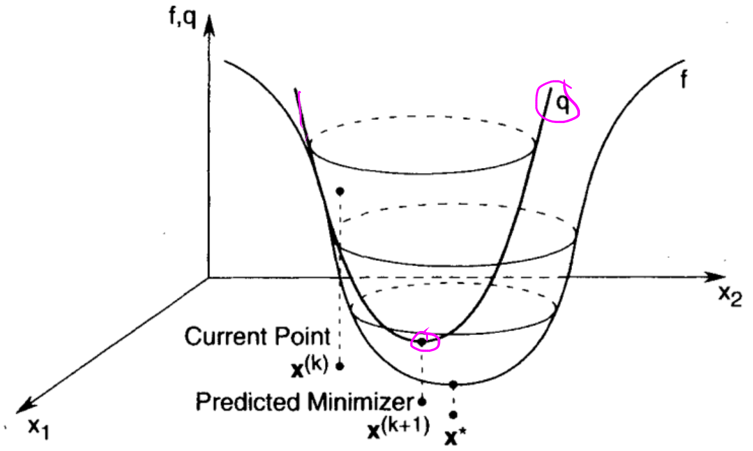
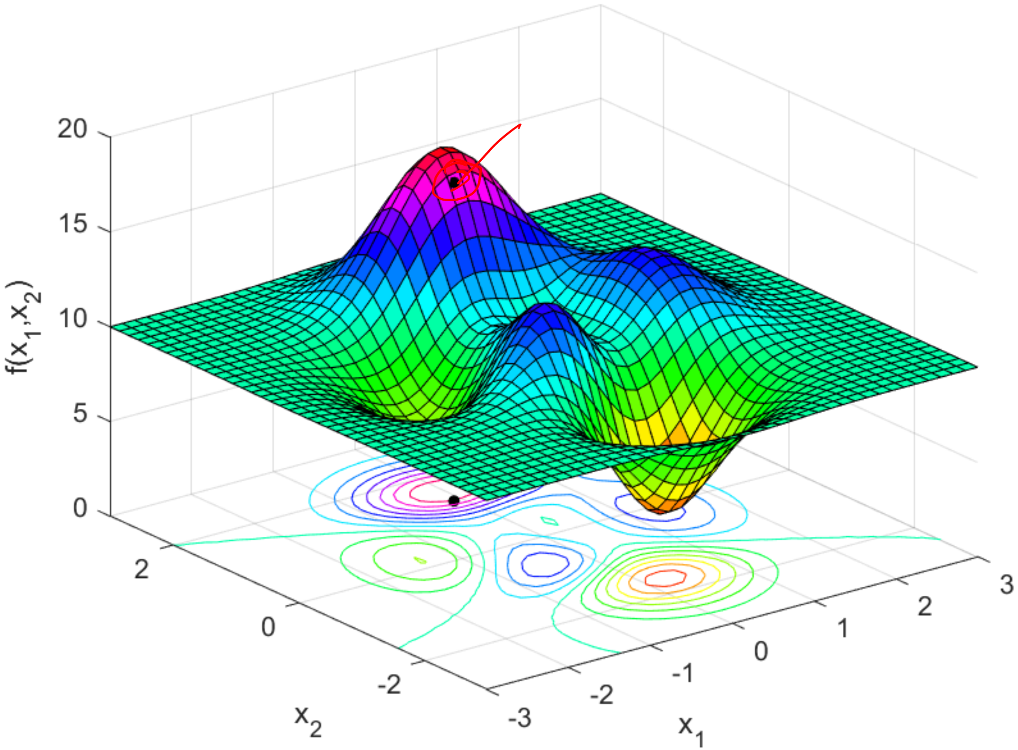


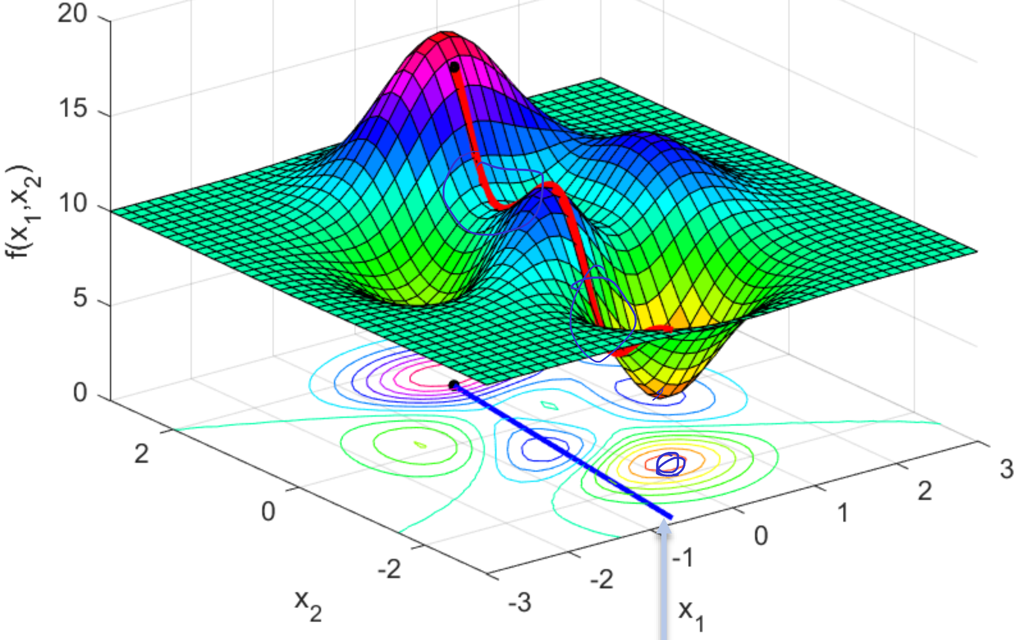
Figure 9.1 Quadratic approximation to the objective function using first and second derivatives.

Chong & Zak, "An introduction to optimization"

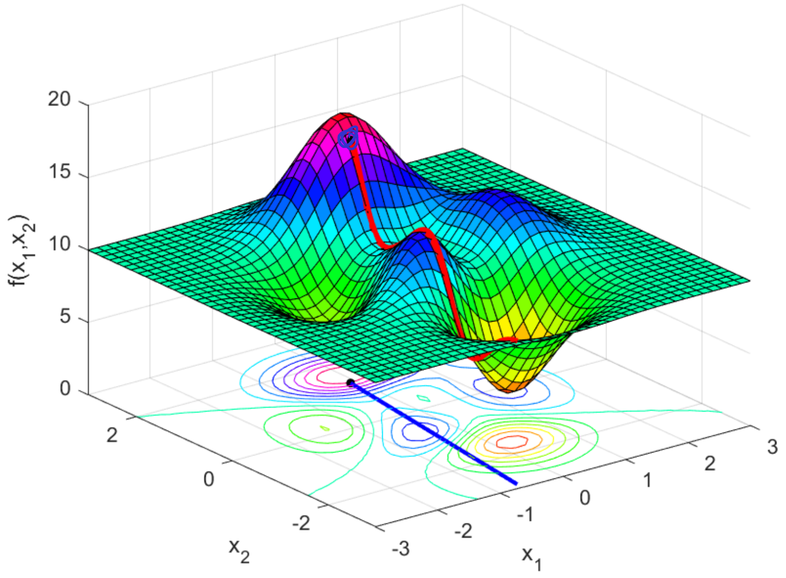
$$x_0 = (-0.1, 1.3)^\top$$



$$x_0 = (-0.1, 1.3)^\top$$

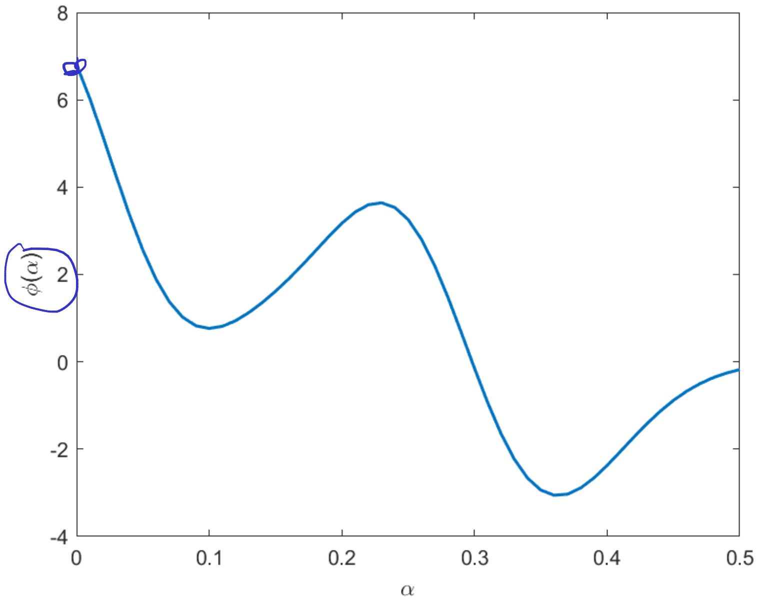


Steepest descent direction from x_0



Define $\phi(\alpha) := f(x_k + \alpha p_k)$

"univariate"



Exact linesearch

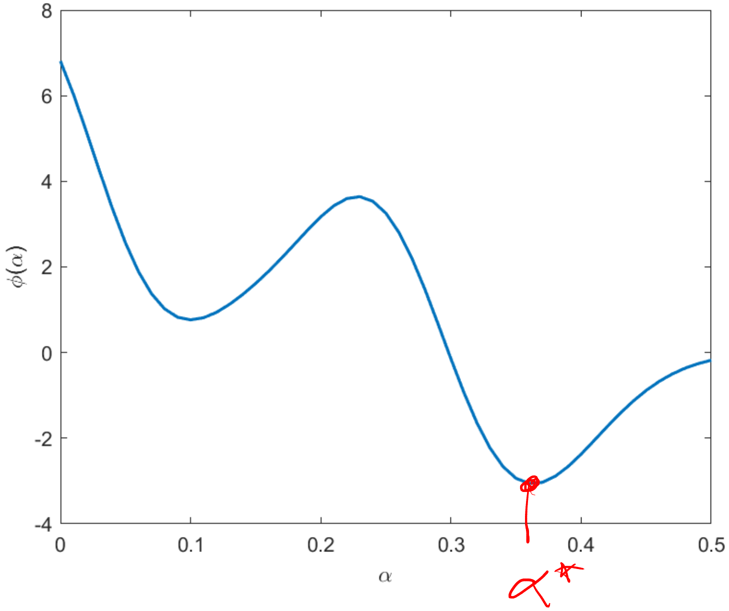
$$\alpha^* = \arg\min_{\alpha} \phi(\alpha)$$

In general, too expensive and unneccessary to compute.

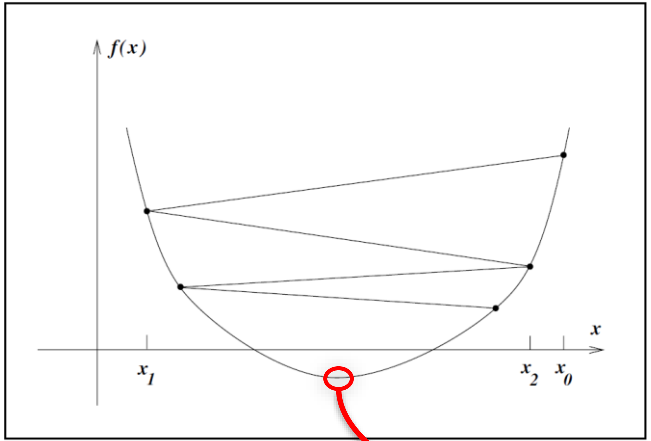
Therefore: Instead find "cheap" α that fulfills

1) Sufficient decrease

2) Desired slope



Why sufficient decrease?



$$f(x_k) = 5/k \to 0 \quad \text{when} \quad k \to \infty$$

$$f(x^*) = -1$$

- Decrease ($f(x_k + \alpha p_k) \leq f(x_k)$) not enough, need <u>sufficient decrease</u> (1st Wolfe condition)

Condition 1: Sufficient decrease (*Armijo*)

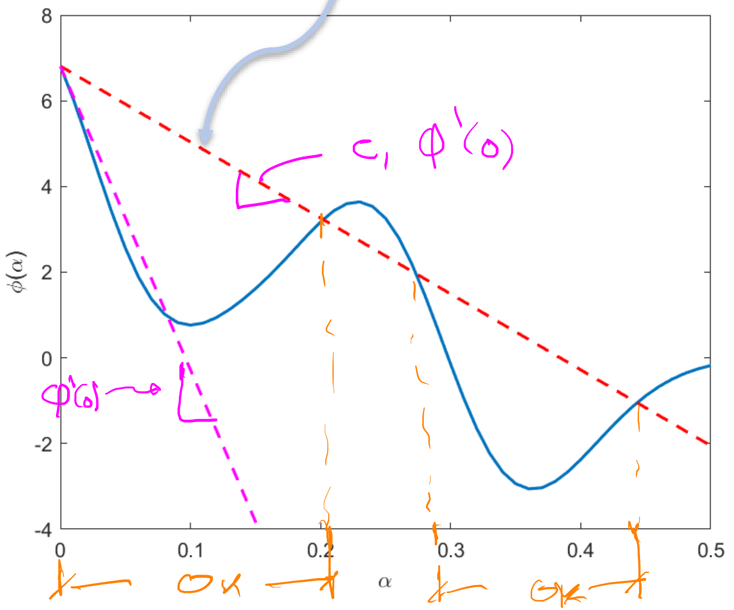
Choose α that fulfills

$$\phi(\alpha) \leq \underbrace{\phi(0) + c_1 \alpha \overbrace{\phi'(0)}^{<0}}_{l(\alpha)}$$

Typical value: $c_1 = 10^{-4}$

Note: Allows (too) small steps.

$l(\alpha) = \phi(0) + c_1 \alpha \phi'(0), \quad c_1 = 0.25$



Sufficient decrease

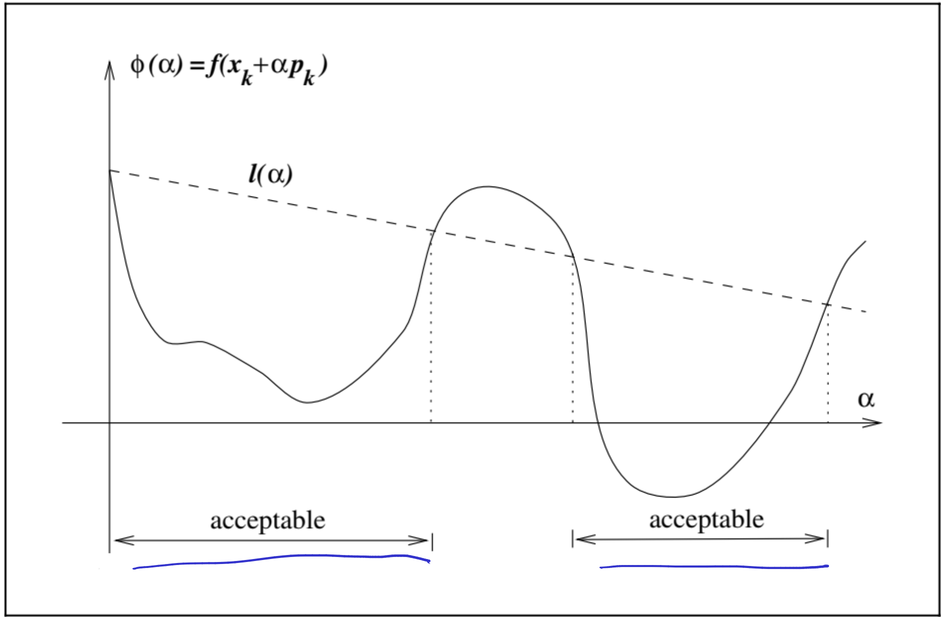


Figure 3.3 Sufficient decrease condition.

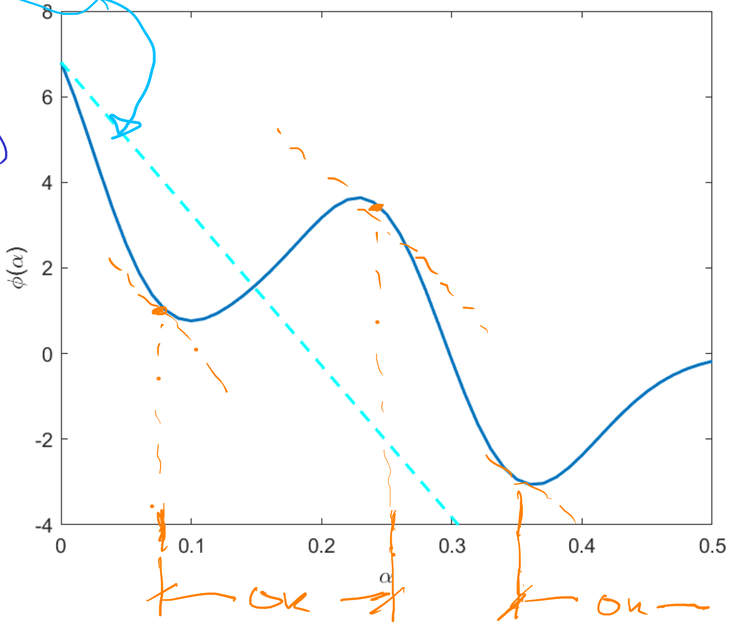
Condition 2: Desired slope

$\phi'(\alpha) > \underline{c_2 \phi'(0)}, \quad c_2 \in (c_1, 1)$

Rationale: If $\phi'(\alpha) \ll 0$ ("steep") don't stop

Typical value: $c_2 = 0.9$

$l(\alpha) = \phi(0) + c_2\alpha\phi'(0), \quad c_2 = 0.5$



Desired slope

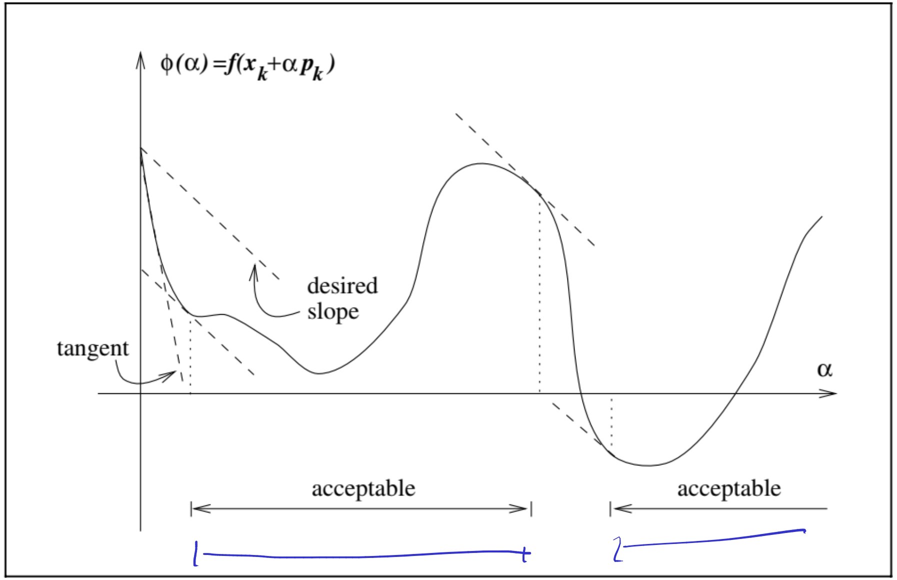


Figure 3.4 The curvature condition.

Wolfe conditions

- Good step lengths should fulfill the Wolfe conditions:

$$\overbrace{f(x_k + \alpha_k p_k)}^{\phi(\alpha_k)} \leq \overbrace{f(x_k)}^{\phi(0)} + c_1 \alpha_k \overbrace{\nabla f_k^\top p_k}^{\phi'(0)} \qquad \text{Sufficient decrease (Armijo condition)}$$

$$\underbrace{\nabla f(x_k + \alpha_k p_k)^\top p_k}_{\phi'(\alpha_k)} \geq c_2 \underbrace{\nabla f_k^\top p_k}_{\phi'(0)} \qquad \text{Desired slope (Curvature condition)}$$

- How do we compute such a step length?

Backtracking Line Search

Algorithm 3.1 (Backtracking Line Search).

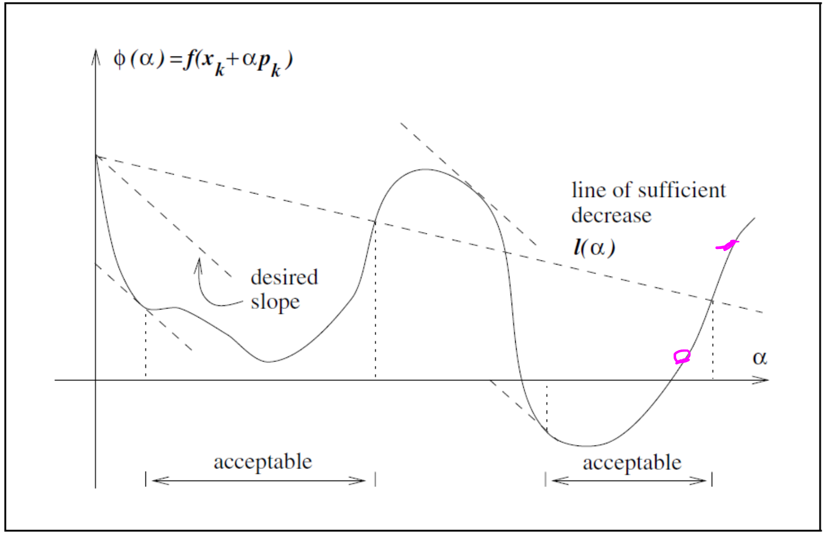
Choose $\bar{\alpha} > 0$, $\rho \in (0, 1)$, $c \in (0, 1)$; Set $\alpha \leftarrow \bar{\alpha}$;

repeat until $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k$

$\alpha \leftarrow \rho\alpha$;

end (repeat)

Terminate with $\alpha_k = \alpha$.



Desired slope (curvature condition) not needed since we start with long step length

Interpolation

① Initialize with α_0 (initial guess)

If α_0 fulfills Wolfe → exit

② Do a quadratic interpolation

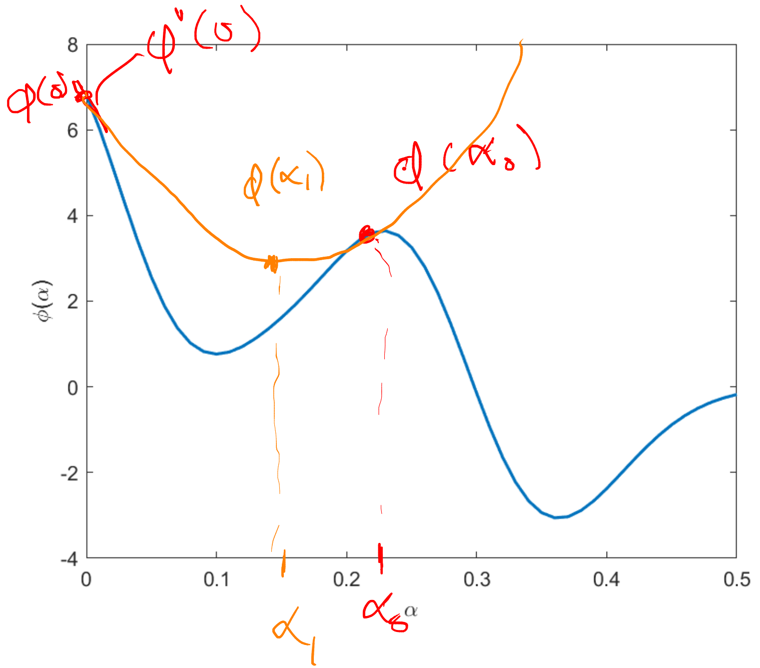
$$\phi_q(\alpha) = a\alpha^2 + \phi'(0)\alpha + \phi(0)$$

$$a = \frac{\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0)}{\alpha_0^2}$$

$$\alpha_1 = \arg\min_{\alpha} \phi_q(\alpha)$$

If α_1 fulfills Wolfe conditions

→ exit



Interpolation

③ Do cubic interpolation.

$$\phi_c(\alpha) = a\alpha^3 + b\alpha^2 + \phi'(0)\alpha + \phi(0)$$

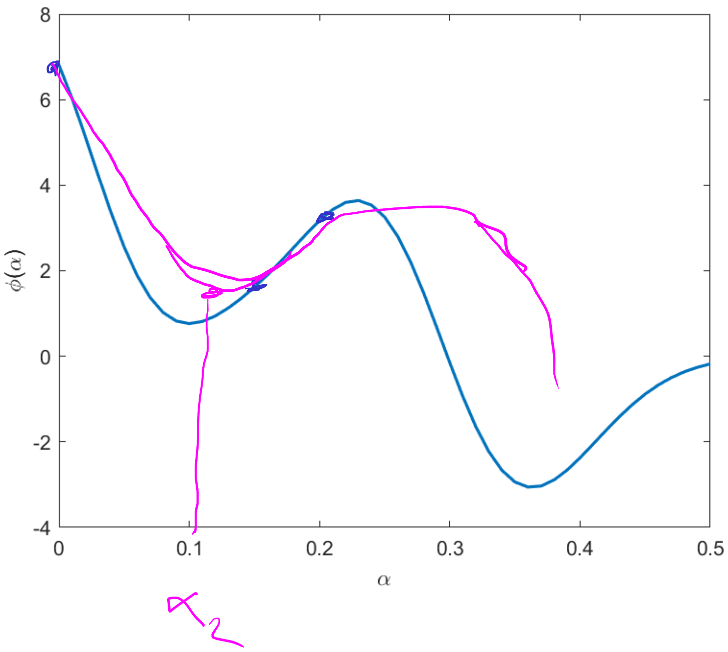
$a = \ldots$ (see book p. 58)

$b = \ldots$

$$\alpha_2 = \arg\min_{\alpha} \phi_c(\alpha) \in [0, \alpha_1]$$

If α_2 fulfills Wolfe $\to$ exit

④ Start over at ① with $\alpha_0 = \alpha_2$



Example: Line search for convex quadratic objective function

$$f(x) = \frac{1}{2}x^\top Gx + c^\top x, \quad G > 0$$

$x_k,\ p_k$ given

$$\phi(\alpha) = \frac{1}{2}(x_k + \alpha p_k)^T G (x_k + \alpha p_k) + c^T (x_k + \alpha p_k)$$

$$= \frac{1}{2} p_k^T G p_k \alpha^2 + x_k^T G p_k \alpha + c^T p_k \alpha + \text{const.}$$

$$\phi'(\alpha) = p_k^T G p_k \cdot \alpha + (x_k^T G + c^T) p_k = 0$$

$$\Rightarrow \boxed{\alpha^* = - \frac{(G x_k + c)^T p_k}{p_k^T G p_k}}$$

If Newton: $p_k = -[\nabla^2 f_k]^{-1} \nabla f_k = -G^{-1}(G x_k + c)$

$$\Rightarrow \alpha^* = + \frac{(G x_k + c)^T G^{-1} (G x_k + c)}{(G x_k + c)^T G^{-1} \cancel{G G^{-1}} (G x_k + c)} = 1$$

Newton: Hessian modification

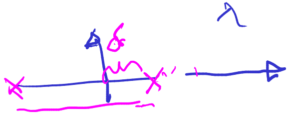
$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k = -(\underline{\nabla^2 f(x_k)})^{-1}\nabla f(x_k)$$ ← In practice: Solve $\boxed{\nabla^2 f_k \, p_k = -\nabla f_k}$

Problem: Far from solution, $\nabla^2 f_k$ typically not pos. def.

Remedy: Replace $\nabla^2 f_k$ with $\nabla^2 f_k + E_k > 0$

E_k can be found in several ways (Ch 3.4)

Ex. $E_k = \tau_k \cdot I, \quad \tau_k = \begin{cases} 0, & \nabla^2 f_k > 0 \\ -\lambda_{min}(\nabla^2 f_k) + \delta, & \text{otherwise} \end{cases}$



Line search Newton

Algorithm 3.2 (Line Search Newton with Modification).

Given initial point x_0;

for $k = 0, 1, 2, \ldots$

Factorize the matrix $B_k = \nabla^2 f(x_k) + E_k$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is sufficiently positive definite; otherwise, E_k is chosen to ensure that B_k is sufficiently positive definite;

Solve $B_k p_k = -\nabla f(x_k)$; ← Cholesky

Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$, where α_k satisfies the Wolfe, Goldstein, or Armijo backtracking conditions;

end

Local convergence rates (close to optimum)

Steepest descent:
Linear convergence

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r \quad \text{for all } k \text{ sufficiently large, } r \in (0,1)$$

Newton:
Quadratic convergence

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq M \quad \text{for all } k \text{ sufficiently large, } M > 0$$

Quasi-Newton:
Superlinear convergence

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

$$\frac{\|x_{k+1} - x^*\|}{\|x_0\|}$$