Lectures 3-4:
Gradient Descent Methods

Optimization T2023

Màster de Fonaments de Ciència de Dades

UNIVERSITAT DE BARCELONA

$$f(\mathbf{x}) \to min, \qquad \mathbf{x} \in D \subseteq \mathbb{R}^n, \quad n \geq 1, \quad f \text{ is smooth}$$

Goal: Iteratively find a sequence $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \to \mathbf{x}^*$, where $\mathbf{x}^*$ is a solution of the optimization problem (local or global minimum), realizing the descent

$$f\left(\mathbf{x}^{(1)}\right) > f\left(\mathbf{x}^{(2)}\right) > \cdots$$

(for all or most* of the iterates)

Recall that $\nabla f(\mathbf{x}^*) = 0$

General descent method.

given a starting point $\mathbf{x}^{(1)} \in D$
repeat

1. Determine *descent direction* $\boldsymbol{p}^{(k)}$ (often, $\|\boldsymbol{p}^{(k)}\| = 1$)
2. Determine *step size/learning rate* $\alpha^{(k)}$
3. Update $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \alpha^{(k)} \boldsymbol{p}^{(k)}$

until *stopping criterion* is satisfied

III. Descent direction?

II. Step size?

I. Stopping criterion?

Digression: Why gradient?

Recall that from the Taylor formula

$$f(\mathbf{x}+\mathbf{v}) =_{up\ to\ H.O.T.} f(\mathbf{x}) + \mathbf{v}^T \cdot \nabla f(\mathbf{x})$$
$$= f(\mathbf{x}) + \nabla_{\mathbf{v}} f(\mathbf{x})$$

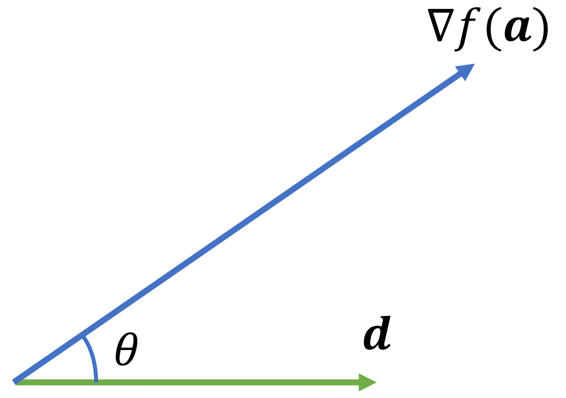
Directional derivative → We want it to be negative

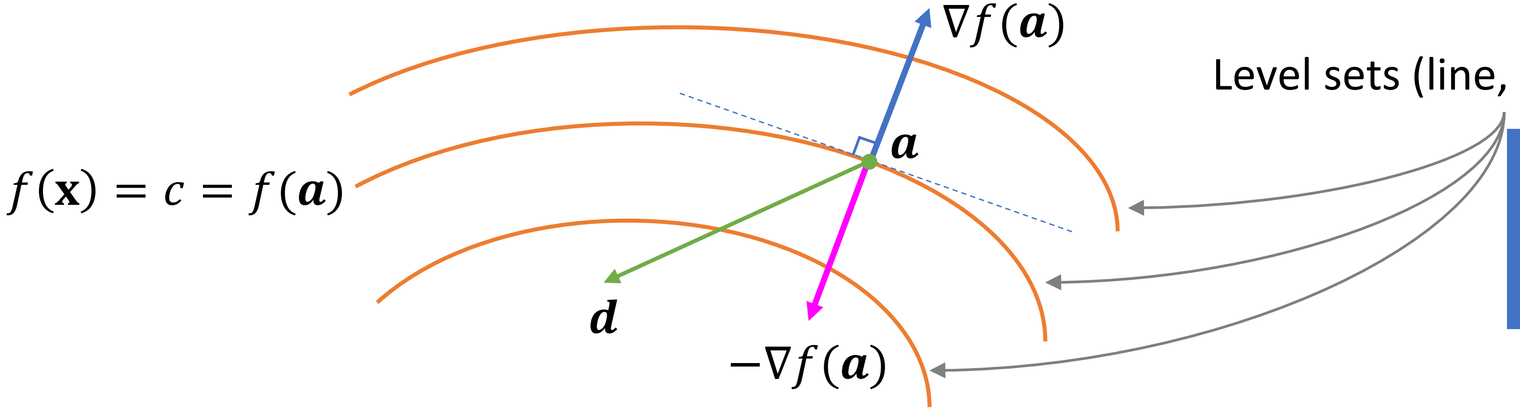
Theorem:

Let $f: D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, $\boldsymbol{a} \in D$, $\boldsymbol{d} \in \mathbb{R}^n$ with $\|\boldsymbol{d}\| = 1$. If θ is the angle between $\boldsymbol{d}$ and $\nabla f(\boldsymbol{a})$. Then

$$\nabla_{\boldsymbol{d}} f(\boldsymbol{a}) = \boldsymbol{d}^T \cdot \nabla f(\boldsymbol{a}) = \|\nabla f(\boldsymbol{a})\| \cos\theta$$

In particular, the vector $-\nabla f(\boldsymbol{a})$ gives the maximum descent direction of f at the point $\boldsymbol{a}$.





$\boldsymbol{d}$ is called a **descent direction** if $\nabla_{\boldsymbol{d}} f(\boldsymbol{a}) = \boldsymbol{d}^T \cdot \nabla f(\boldsymbol{a}) < 0$

I. Stopping criteria/termination conditions

- **Maximum iterations**: repeat until $k \leq k_{max}$
- **Absolute improvement**: repeat until

$$f\left(\mathbf{x}^{(k)}\right) - f\left(\mathbf{x}^{(k+1)}\right) < \epsilon_a$$

- **Relative improvement**: repeat until

$$f\left(\mathbf{x}^{(k)}\right) - f\left(\mathbf{x}^{(k+1)}\right) < \epsilon_r \left| f\left(\mathbf{x}^{(k)}\right) \right|$$

- **Gradient magnitude**: repeat until

$$\left\| \nabla f\left(\mathbf{x}^{(k+1)}\right) \right\| < \epsilon_g$$

✓ One or more termination conditions can be used

✓ If there are several local minima, one can add ***random restart*** with $\mathbf{x}^{(1),new}$ sampled randomly from D

II. Step size/learning rate

Suppose $x = x^{(k)}$ and $p = p^{(k)}$ is given. How to find $\alpha = \alpha^{(k)}$?

Methods:

1. Exact line search
2. Approximate line search
3. Trust region methods

Exact line search

$$\text{minimize}_{\alpha} \; f(\mathbf{x} + \alpha \boldsymbol{p})$$

Definition: A *bracket* is an interval $[\alpha', \alpha"]$ containing α^*

- This is univariant optimization problem for $\phi(\alpha) := f(\mathbf{x} + \alpha \boldsymbol{p})$ $\rightarrow$

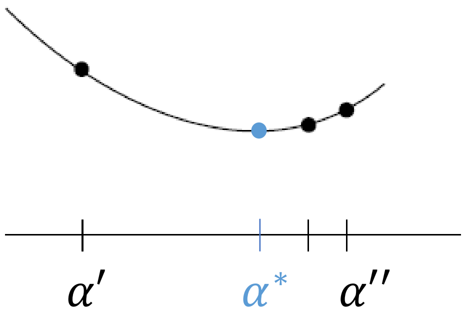
  → Find a **bracket** for the optimal solution α^*

  (α^* is characterized by $\phi(\alpha^*) < \phi(\alpha)$ for all α near α^*)

  → Use univariant optimization methods to find an approximation of α^* by successively shrinking the bracket. Methods include:

  - Dyadic/binary search
  - Fibonacci search
  - Quadratic fit search
  - Shubert–Piyavskii method
  - Bisection method

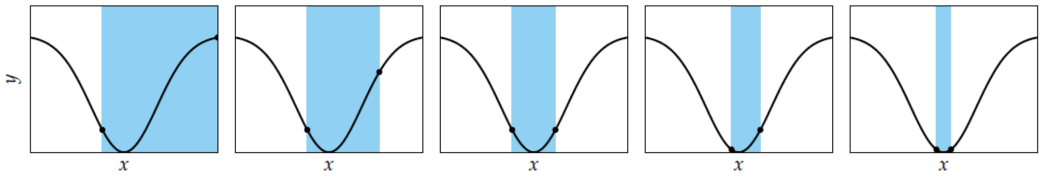
Only for unimodal functions!



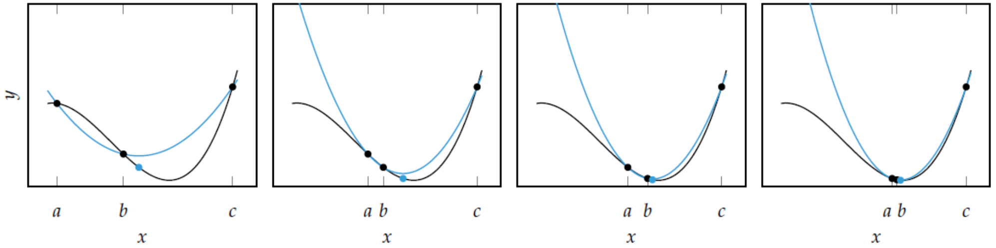
Digression: some univalent optimization methods

[KW, Ch.3]

- **Dyadic/binary search**: subdivide interval 'in half' at each step

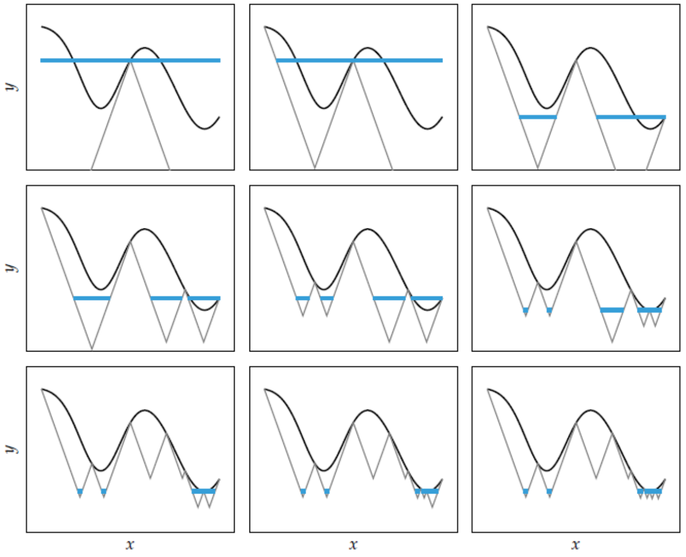


- **Fibonacci search :** max reduction of interval size for given number of function evaluations
- **Quadratic fit search**

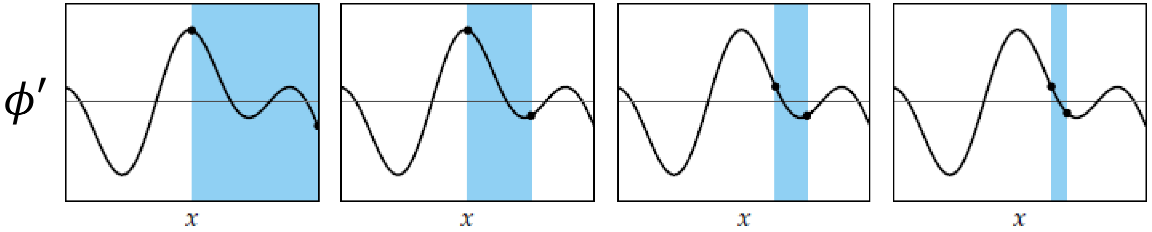


- **Shubert-Piyavskii method** : assuming ϕ is Lipshitz, e.g.

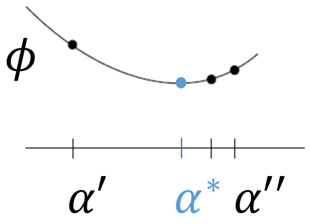
$$|\phi(x) - \phi(y)| \leq \ell \cdot |x - y|, \quad \forall x, y \in [\alpha', \alpha'']$$



- **Bisection method**: solve $\phi'(\alpha) = 0$ instead



Dyadic/binary and Fibonacci search



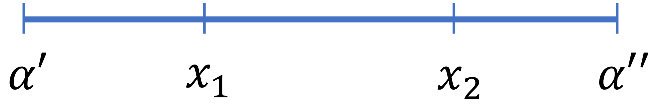
Assumption ★:

ϕ is unimodal, that is, ϕ has a unique minimum on (α', α'')

$\Rightarrow$ ϕ is decreasing on $[\alpha', \alpha^*]$ and ϕ is increasing on $[\alpha^*, \alpha'']$

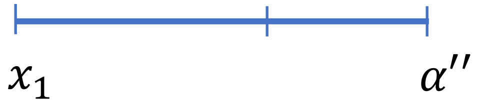
$\Rightarrow$ ϕ is convex on $[\alpha', \alpha'']$ ($\Leftrightarrow \phi'' > 0$)

Basic Splitting Step: for a pair $x_1 < x_2$ of points in the starting bracket



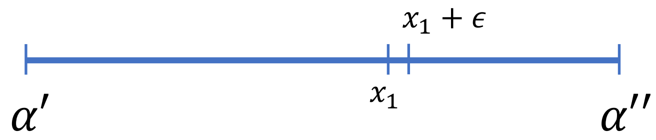
New interval if $\phi(x_2) > \phi(x_1)$

New interval if $\phi(x_2) < \phi(x_1)$



Exercise: Check that, under Assumption ★, for all $x_1 < x_2$ after the Basic Splitting Step the new interval contains α^* (hence, is a bracket).

Basic Splitting Step in "almost" two parts: do the basic splitting step for x_1 and $x_1 + \epsilon$, where $\epsilon > 0$ is small



- Each Basic Splitting Step requires 2 evaluations of the function at x_1 and x_2.
- In general, i.e., if Assumption ★ is violated, the Basic Splitting Step doesn't work!

Exercise: Give an example

Dyadic/binary search:
under Assumption ★

given the desired size $\epsilon > 0$ of the bracket

choose $\delta < \epsilon$ (usually much smaller)

repeat

1. Pick the midpoint $x_1 = \frac{\alpha' + \alpha''}{2}$
2. Do the **Basic Splitting Step in 'almost' two parts** using x_1 and $x_1 + \delta$
3. Update $[\alpha', \alpha'']$ with the new bracket from step 2 above

until $|\alpha'' - \alpha'| < \epsilon$

Exercise: How many evaluations of the function ϕ is required in the dyadic search in order to shrink the bracket by a factor of 100?

Fibonacci search (under Assumption ★)

- Fibonacci numbers are given by the recursive relation $F_{n+2} = F_{n+1} + F_n$, with starting condition $F_1 = F_2 = 1$.
- This generates the sequence $1, 1, 2, 3, 5, 8, 11, \ldots$
- This sequence grows as $F_n \sim \frac{1}{\sqrt{5}}\varphi^n$, for n large, where $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.6$ is the Golden ratio.

---

given the number of steps N

for $i = N, N-1, \ldots, 1$ **do**

if $i \neq 1$,

1. Compute $x_1, x_2 \in [\alpha', \alpha'']$ such that

$$\frac{\alpha''-x_1}{\alpha''-\alpha'} = \frac{F_i}{F_{i+1}} \text{ and } \frac{x_2-\alpha'}{\alpha''-\alpha'} = \frac{F_i}{F_{i+1}}$$

2. Do the **Basic Splitting Step** using x_1 and x_2
3. Update $[\alpha', \alpha'']$ with the new bracket from step 2 above

Observe that after this step, the length of the new bracket is proportional to the length of the previous bracket as F_i to F_{i+1}

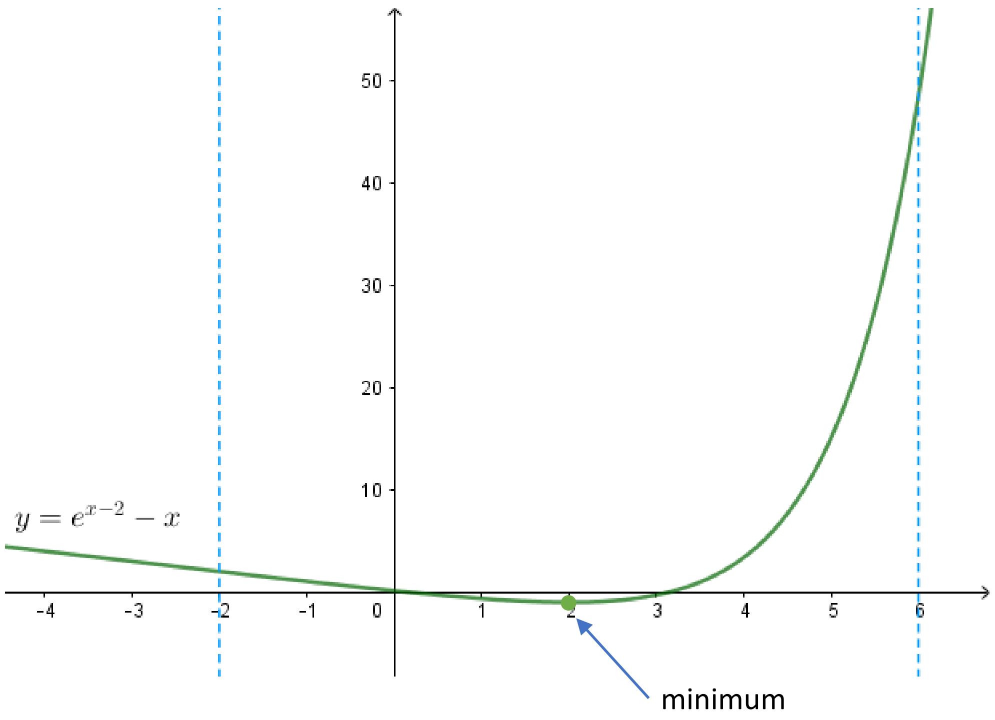
otherwise

Do the **Basic Splitting Step in 'almost' two parts** using $\frac{\alpha'+\alpha''}{2}$ and $\frac{\alpha'+\alpha''}{2} + \epsilon$

Key Advantage: Fibonacci search uses significantly smaller evaluations of the function than the dyadic search because it re-uses some evaluation points! (see example on the next slide)

Exercise: How many evaluations of the function ϕ is required in the Fibonacci search in order to shrink the bracket by a factor of 100? Compare it to the corresponding result of the dyadic search.

Fibonacci search (under Assumption ★): an example



Consider using Fibonacci search with five function evaluations to minimize $f(x) = \exp(x-2) - x$ over the interval $[a, b] = [-2, 6]$. The first two function evaluations are made at $\frac{F_5}{F_6}$ and $1 - \frac{F_5}{F_6}$, along the length of the initial bracketing interval:

$$f(x^{(1)}) = f\left(a + (b-a)\left(1 - \frac{F_5}{F_6}\right)\right) \qquad = f(1) = -0.632$$
$$f(x^{(2)}) = f\left(a + (b-a)\frac{F_5}{F_6}\right) \qquad = f(3) = -0.282$$

The evaluation at $x^{(1)}$ is lower, yielding the new interval $[a, b] = [-2, 3]$. Two evaluations are needed for the next interval split:

$$x_{\text{left}} = a + (b-a)\left(1 - \frac{F_4}{F_5}\right) = 0$$
$$x_{\text{right}} = a + (b-a)\frac{F_4}{F_5} = 1$$

A third function evaluation is thus made at x_{left}, as x_{right} has already been evaluated:

$$f(x^{(3)}) = f(0) = 0.135$$

The evaluation at $x^{(1)}$ is lower, yielding the new interval $[a, b] = [0, 3]$. Two evaluations are needed for the next interval split:

$$x_{\text{left}} = a + (b-a)\left(1 - \frac{F_3}{F_4}\right) = 1$$
$$x_{\text{right}} = a + (b-a)\frac{F_3}{F_4} = 2$$

A fourth functional evaluation is thus made at x_{right}, as x_{left} has already been evaluated:

$$f(x^{(4)}) = f(2) = -1$$

The new interval is $[a, b] = [1, 3]$. A final evaluation is made just next to the center of the interval at $2 + \epsilon$, and it is found to have a slightly higher value than $f(2)$. The final interval is $[1, 2 + \epsilon]$.

Quadratic fit search

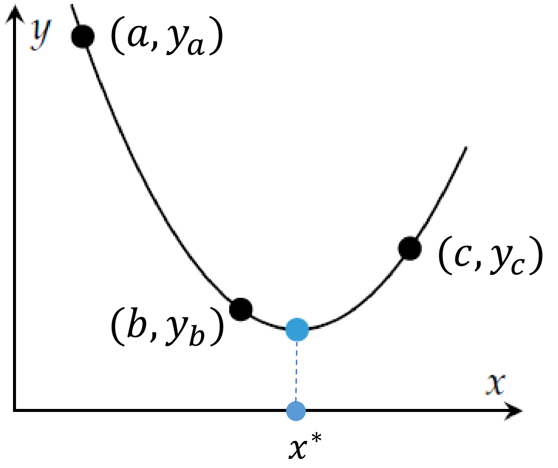
The method is based on the following observations:

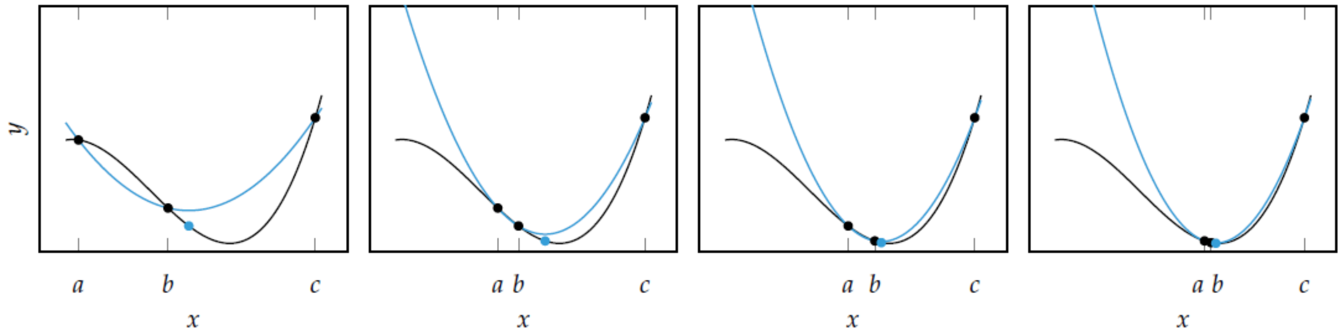
- ‘close’ to the minima functions look like quadratic functions
- we can explicitly find minima of quadratic functions:

Lemma:

There exists a unique parabola that passes through any triple of distinct points $(a, y_a), (b, y_b), (c, y_c)$. This parabola has its extremum at

$$x^* = \frac{1}{2}\frac{y_a(b^2 - c^2) + y_b(c^2 - a^2) + y_c(a^2 - b^2)}{y_a(b - c) + y_b(c - a) + y_c(a - b)}$$





Exercise*: Show that the algorithm described on the next slide converges to a local minimum (assuming the function is smooth)

Quadratic fit search

given a triple $a < b < c$ where $[a, c]$ is a bracket of ϕ and $\overbrace{\phi(b) < \phi(a),\ \phi(b) < \phi(c)}^{\#}$
repeat

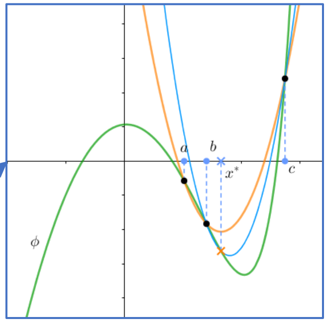
1. Compute the critical point x^* of the parabola passing through a, b, c
2. If $x^* \in [b, c]$, then
   - Check which value is larger, $\phi(x^*)$ or $\phi(b)$:
     i. If $\phi(b) > \phi(x^*)$, update the triple (a, b, c) with (b, x^*, c)
     ii. If $\phi(b) < \phi(x^*)$, update the triple (a, b, c) with (a, b, x^*)
3. Otherwise
   - Again, check which value is larger, $\phi(x^*)$ or $\phi(b)$:
     i. If $\phi(b) > \phi(x^*)$, update the triple (a, b, c) with (a, x^*, b)
     ii. If $\phi(b) < \phi(x^*)$, update the triple (a, b, c) with (x^*, b, c)

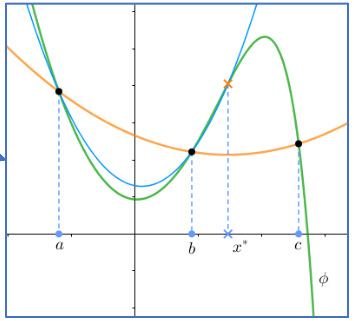
until the $|a - c| < \epsilon$

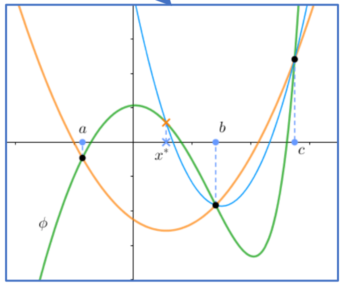
That is, when $x^* \in [a, b)$ because of condition #

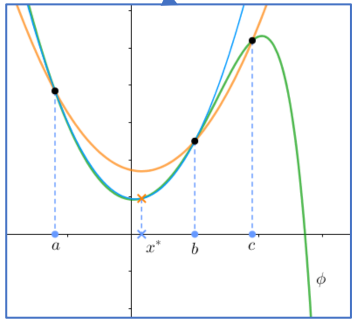
Or any other stopping criterion based on variation of the function

In these examples, parabola at the current step is in orange; parabola at the next step is in blue









Bisection method

The method is based on the following observations:

- Instead of looking for a local minimum of ϕ, we can look for a solution of $\phi' = 0$
- We assume that $[\alpha', \alpha'']$ is a bracket for ϕ, and hence there exists a solution of $\phi' = 0$ on this interval

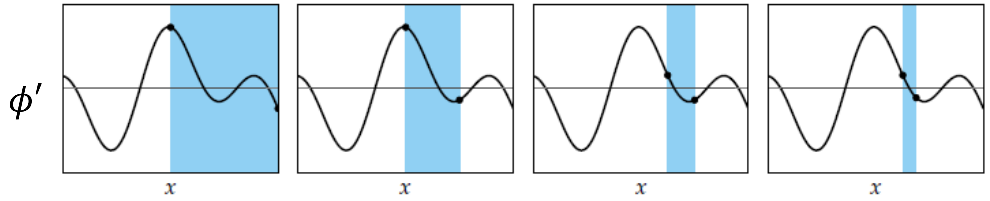
---

given an interval $[a, b]$ such that $\phi'(a) \cdot \phi'(b) < 0$

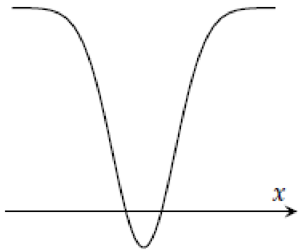
repeat

1. Compute the midpoint $c = \frac{a+b}{2}$
2. If $\phi'(a) \cdot \phi'(c) < 0$, update interval $[a, b]$ with $[a, c]$
3. If $\phi'(b) \cdot \phi'(c) < 0$, update interval $[a, b]$ with $[c, b]$

until $|a - b| < \epsilon$



- If $[\alpha', \alpha'']$ doesn't satisfy the condition $\phi'(\alpha') \cdot \phi'(\alpha'') < 0$, then one can try iteratively shrink this interval by a constant factor (say 2), until the condition is fulfilled. However, it might not always work (see an example of the function on the left where the bisection method can fail; this is the situation of a local minimum in a 'deep valley'). More sophisticated methods should be used instead.



Exercise: Let $\phi(x) = \frac{x^2}{2} - x$. Apply the bisection method to find an interval containing the minimizer of ϕ starting with the interval $[0,1000]$. Execute 3 steps of the algorithm.

End of digression

Exact line search

$$\text{minimize}_\alpha \; f(\mathbf{x} + \alpha \boldsymbol{p})$$

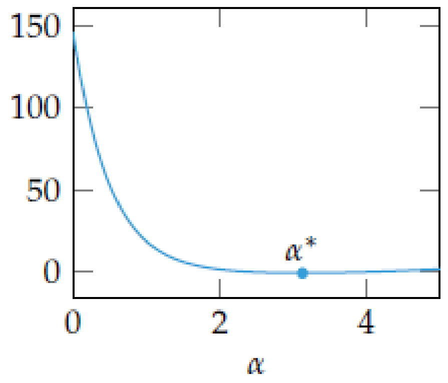
Consider conducting a line search on $f(x_1, x_2, x_3) = \sin(x_1 x_2) + \exp(x_2 + x_3) - x_3$ from $\mathbf{x} = [1, 2, 3]$ in the direction $\mathbf{d} = [0, -1, -1]$. The corresponding optimization problem is:

$$\underset{\alpha}{\text{minimize}} \sin((1 + 0\alpha)(2 - \alpha)) + \exp((2 - \alpha) + (3 - \alpha)) - (3 - \alpha)$$

which simplifies to:

$$\underset{\alpha}{\text{minimize}} \sin(2 - \alpha) + \exp(5 - 2\alpha) + \alpha - 3$$

The minimum is at $\alpha \approx 3.127$ with $\mathbf{x} \approx [1, -1.126, -0.126]$.



Approximate line search

Find $\alpha^{(k)}$ so that the value $f(\mathbf{x}_k + \alpha^{(k)}\boldsymbol{p}^{(k)})$ decreases (not necessarily best possible) and move on with the descent method

For simplicity, $x_k = \mathbf{x}^{(k)}, p_k = \boldsymbol{p}^{(k)}, \alpha_k = \alpha^{(k)}$

We impose the following condition for α_k

$$\phi(\alpha_k) := f(x_k + \alpha_k p_k) < f(x_k) + c_1 \alpha_k (\nabla f(x_k))^T p_k, \quad c_1 \in (0,1).$$

The condition is called (sufficient decrease condition).

Remarks.

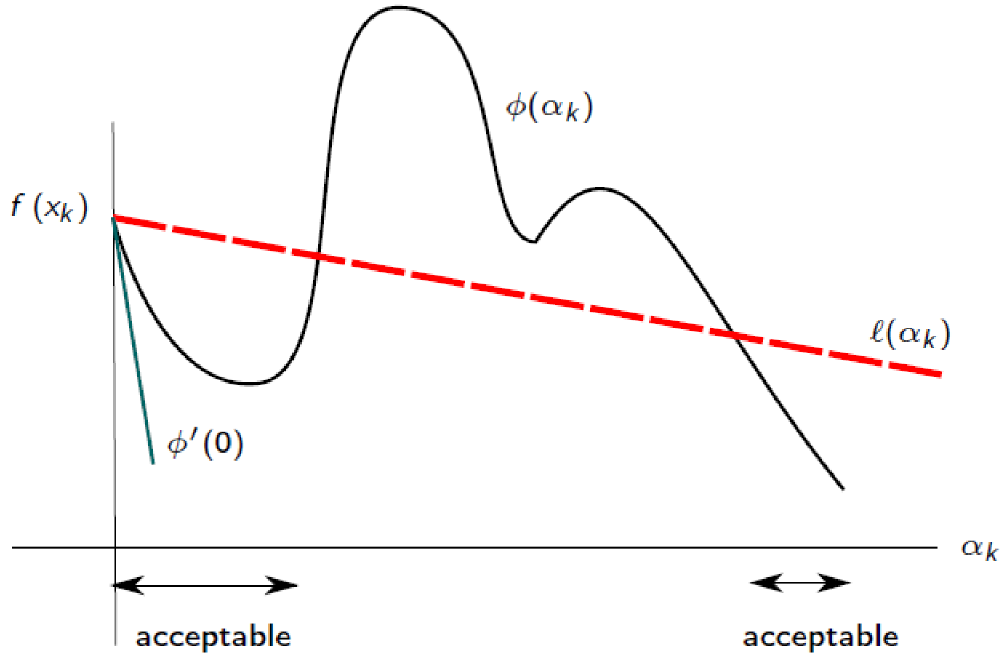
- $\ell(\alpha_k) := f(x_k) + c_1 \alpha_k \nabla f^T(x_k) p_k$ is a linear function.
- For small values of $\alpha_k > 0$ we have $\phi(\alpha_k) < \ell(\alpha_k)$. This is so because $c_1 \in (0,1)$ and then

$$\phi'(0) = (\nabla f(x_k))^T p_k < c_1 (\nabla f(x_k))^T p_k = \ell'(0) < 0.$$

Recall: since p_k is a descent direction, we have $(\nabla f(x_k))^T p_k < 0$

Approximate line search

Sufficient decrease. We ask for a decrease proportional to α and $\phi'(0) = (\nabla f(x_k))^T p_k$. Usually $c_1 \approx 0.1$.

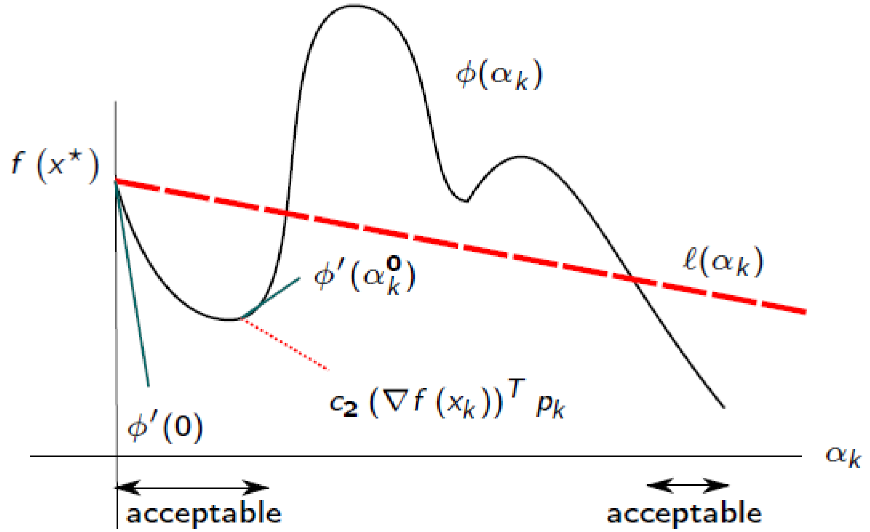


Approximate line search

Curvature condition. Since the previous condition is always satisfied for small values of α_k we need to add further conditions for termination. We use the so called curvature condition

$$(\nabla f(x_k + \alpha_k p_k))^T p_k \geq c_2 (\nabla f(x_k))^T p_k, \;\; c_2 \in (c_1, 1)$$

In other words if $\phi'(\alpha_k)$ is not negative enough we terminate the k-step.



Approximate line search

Wolfe conditions

Definition. The conditions (together) to terminate the k-step given by

$$f\left(x_k+\alpha_k p_k\right)<f\left(x_k\right)+c_1\alpha_k\left(\nabla f\left(x_k\right)\right)^T p_k,$$
$$\left(\nabla f\left(x_k+\alpha_k p_k\right)\right)^T p_k \geq c_2\left(\nabla f\left(x_k\right)\right)^T p_k,$$

with $0<c_1<c_2<1$ are usually called Wolfe conditions.

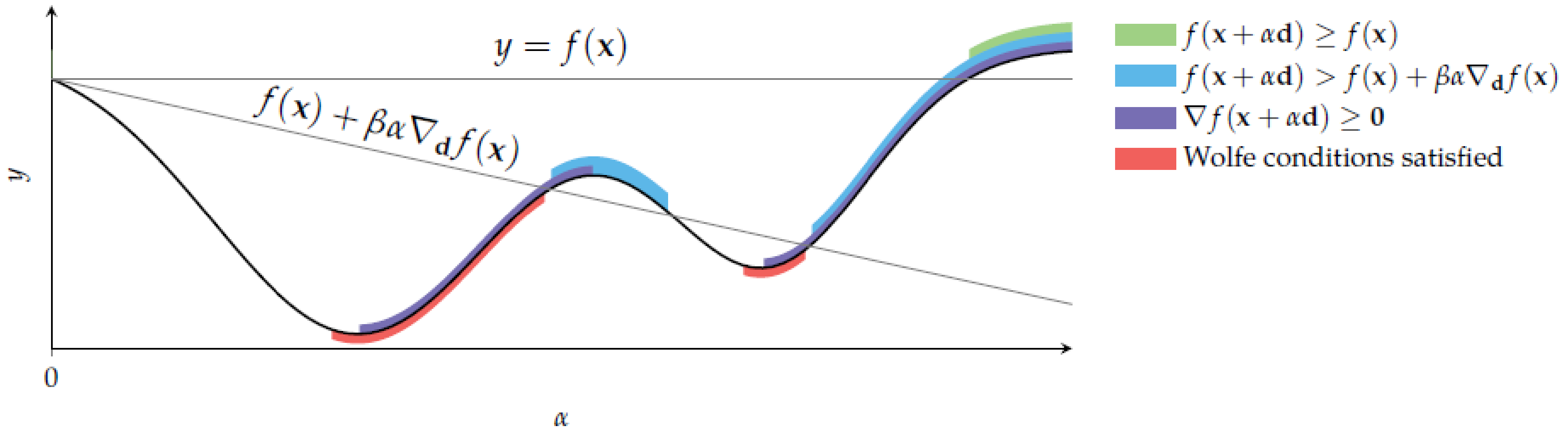
Definition. The conditions (together) to terminate the k-step given by (we do not allow $\phi'\left(\alpha_k\right)$ to be too positive).

$$f\left(x_k+\alpha_k p_k\right)<f\left(x_k\right)+c_1\alpha_k\left(\nabla f\left(x_k\right)\right)^T p_k,$$
$$\left|\left(\nabla f\left(x_k+\alpha_k p_k\right)\right)^T p_k\right| \leq \left|c_2\left(\nabla f\left(x_k\right)\right)^T p_k\right|,$$

with $0<c_1<c_2<1$ are usually called strong Wolfe conditions.

Approximate line search

Wolfe conditions



Approximate line search

Wolfe conditions: existence

Lemma. Suppose $f : D \subset \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$ function. Let p_k a descent direction at the point $x_k \in D$ and assume $f|L_{p_k}$ is bounded below where $L_{p_k} = \{x \in \mathbb{R}^n \mid x = x_k + \alpha p_k,\ \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$ there exist intervals of step lengths satisfying the (strong) Wolfe conditions

Proof. Since $\ell'(\alpha_k) < 0$ (and constant) there exists a first intersection, $\hat{\alpha}_k > 0$, between $\ell(\alpha_k)$ and $\phi(\alpha_k)$:

$$f(x_k + \hat{\alpha}_k p_k) = f(x_k) + c_1 \hat{\alpha}_k (\nabla f(x_k))^T p_k. \tag{1}$$

The sufficient decrease condition it is satisfied for all $\alpha_k \in [0, \hat{\alpha}_k]$. By the Mean Value Theorem we have that there exists $\tilde{\alpha}_k \in [0, \hat{\alpha}_k]$ such that

$$f(x_k + \hat{\alpha}_k p_k) - f(x_k) = \hat{\alpha}_k (\nabla f(x_k + \tilde{\alpha}_k p_k))^T p_k$$

All together imply

$$(\nabla f(x_k + \tilde{\alpha}_k p_k))^T p_k = c_1 \hat{\alpha}_k (\nabla f(x_k))^T p_k > c_2 \hat{\alpha}_k (\nabla f(x_k))^T p_k.$$

Therefore $\tilde{\alpha}_k$ satisfies the Wolfe conditions and smoothness gives the desired interval.

Approximate line search

Convergence

Remark. Until this moment we just consider the definition of the process, that is the election of p_k and α_k. But we need to study if the process converge to somewhere.

Let p_k be a descent direction, and let θ_k the angle of p_k and $-\nabla f(x^\star)$

$$\cos(\theta_k) = -\frac{1}{||\nabla f(x_k)|| \, ||p_k||} (\nabla f(x_k))^T p_k$$

Theorem. Assume notation above with p_k a descent direction and α_k satisfying Wolfe's conditions. Suppose f is $\mathcal{C}^2$ and bounded below in $\mathbb{R}^n$. Then

$$\sum_{k=0}^{\infty} \cos^2(\theta_k) ||\nabla f(x_k)|| < \infty. \tag{2}$$

Approximate line search | Convergence

Corollary. Under the above notation and assumptions we have

$$\cos^2(\theta_k)||\nabla f(x_k)|| \to 0$$

Moreover if there exists $\delta > 0$ such that $\cos(\theta) > \delta$ then

$$\lim_{k\to\infty} ||\nabla f(x_k)|| = 0 \quad \text{(globally convergent algorithms)}$$

Remark. The final δ-condition basically means that p_k do not get arbitrarily orthogonal to the gradient vector. This is, for instance, the case of the steepest descent method.

Approximate line search **Convergence**

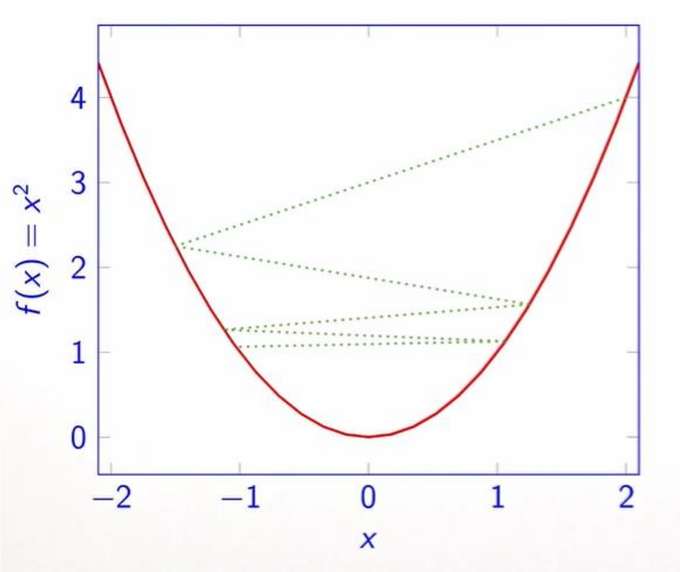
Exercise: Consider the function $f(x) = x^2$ on $[-2,2]$. Consider the one-dimensional gradient descent method starting at $\mathbf{x}_0 = 2$ in the direction

$$\boldsymbol{p}_k = -sign(\mathbf{x}_k)$$

with step

$$\alpha_k = 2 + 3(2^{-k-1}).$$

1) Verify that $\boldsymbol{p}_k$ is indeed a descent direction, that is, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.
2) Perform 5 steps of the descent algorithm.
3) Does this descent converge? (*Hint: see picture on the right.*) Justify your argument. What Wolfe conditions are violated?



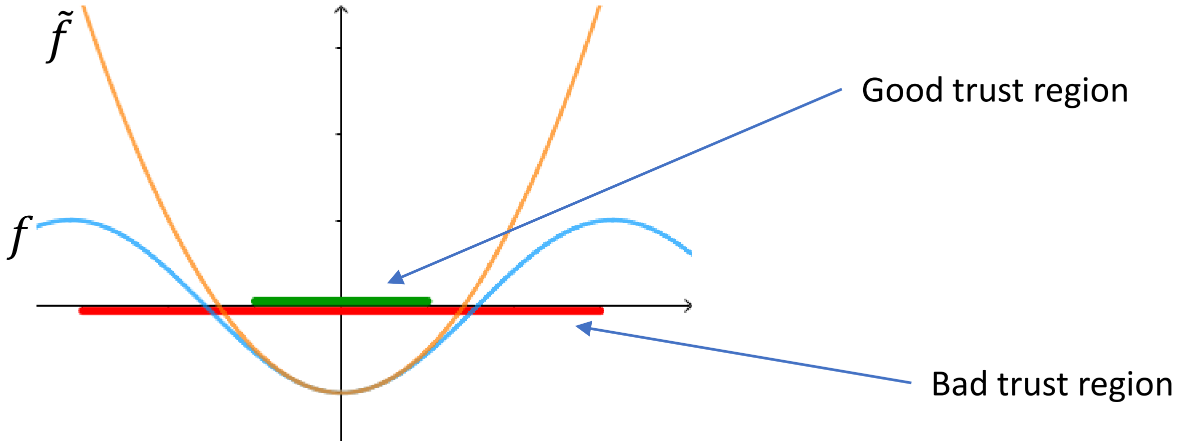
Trust region method

Idea

- Line search methods: find a descent direction → find the next point in this direction

- Trust region methods: find a region 'of possible good steps' → find a point in this region

Usually, we approximate the objective function f with a simpler objective $\tilde{f}$.



Potential problem: It might be that the solution $\tilde{x}^*$of $\min \tilde{f}(x)$ lies in the region where $\tilde{f}$ badly approximate f

A solution: restrict the optimization of $\tilde{f}$ to the region where we **trust** that $\tilde{f}$ is a good approximation of f

Trust region method

Idea (cont.)

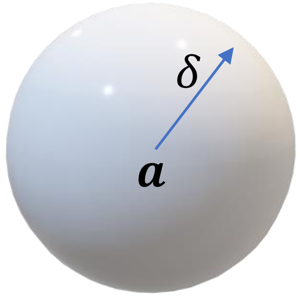
Typically, near a point $\boldsymbol{a}$ we do the quadratic approximation

$$f(\mathbf{x}) \approx \tilde{f}(\mathbf{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T \cdot (\mathbf{x} - \boldsymbol{a}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{a})^T \cdot \nabla^2 f(\boldsymbol{a}) \cdot (\mathbf{x} - \boldsymbol{a})$$

At $\boldsymbol{a}$, f and $\tilde{f}$ match: $f(\boldsymbol{a}) = \tilde{f}(\boldsymbol{a})$
The further we go from $\boldsymbol{a}$, the worse is the approximation

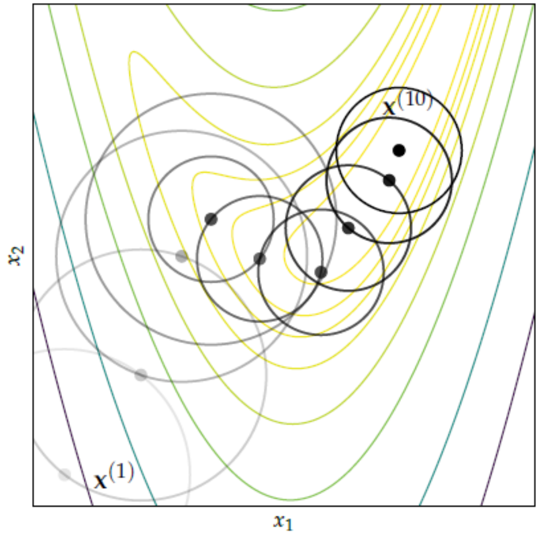
A trust region might be a ball of radius $\delta > 0$ centered at $\boldsymbol{a}$:

$$\{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \boldsymbol{a}\| \leq \delta\}$$



Trust region method

Generic algorithm



given $\delta, \mathbf{x}_1$ and $k = 0$
repeat

1. $k \leftarrow k + 1$
2. Find a solution $\mathbf{x}_k^*$ of the minimization problem $\tilde{f} \to min$ subject to $\|\mathbf{x} - \mathbf{x}_{k-1}^*\| \leq \delta$
3. If $\tilde{f}(\mathbf{x}_k^*) \approx f(\mathbf{x}_k^*)$, then increase δ
   else descrease δ

until the required precision is reached

Trust region subproblem

For example, for

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}_{k-1}^*) + \nabla f(\mathbf{x}_{k-1}^*)^T \cdot (\mathbf{x} - \mathbf{x}_{k-1}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{k-1}^*)^T \cdot \nabla^2 f(\mathbf{x}_{k-1}^*) \cdot (\mathbf{x} - \mathbf{x}_{k-1}^*)$$

For example: compute the *predictive performance*

$$\eta = \frac{\text{actual improvement}}{\text{predicted improvment}} = \frac{f(x_{k-1}^*) - f(x_k^*)}{f(x_{k-1}^*) - \tilde{f}(x_k^*)} \in (0,1]$$

$= \tilde{f}(x_{k-1}^*)$

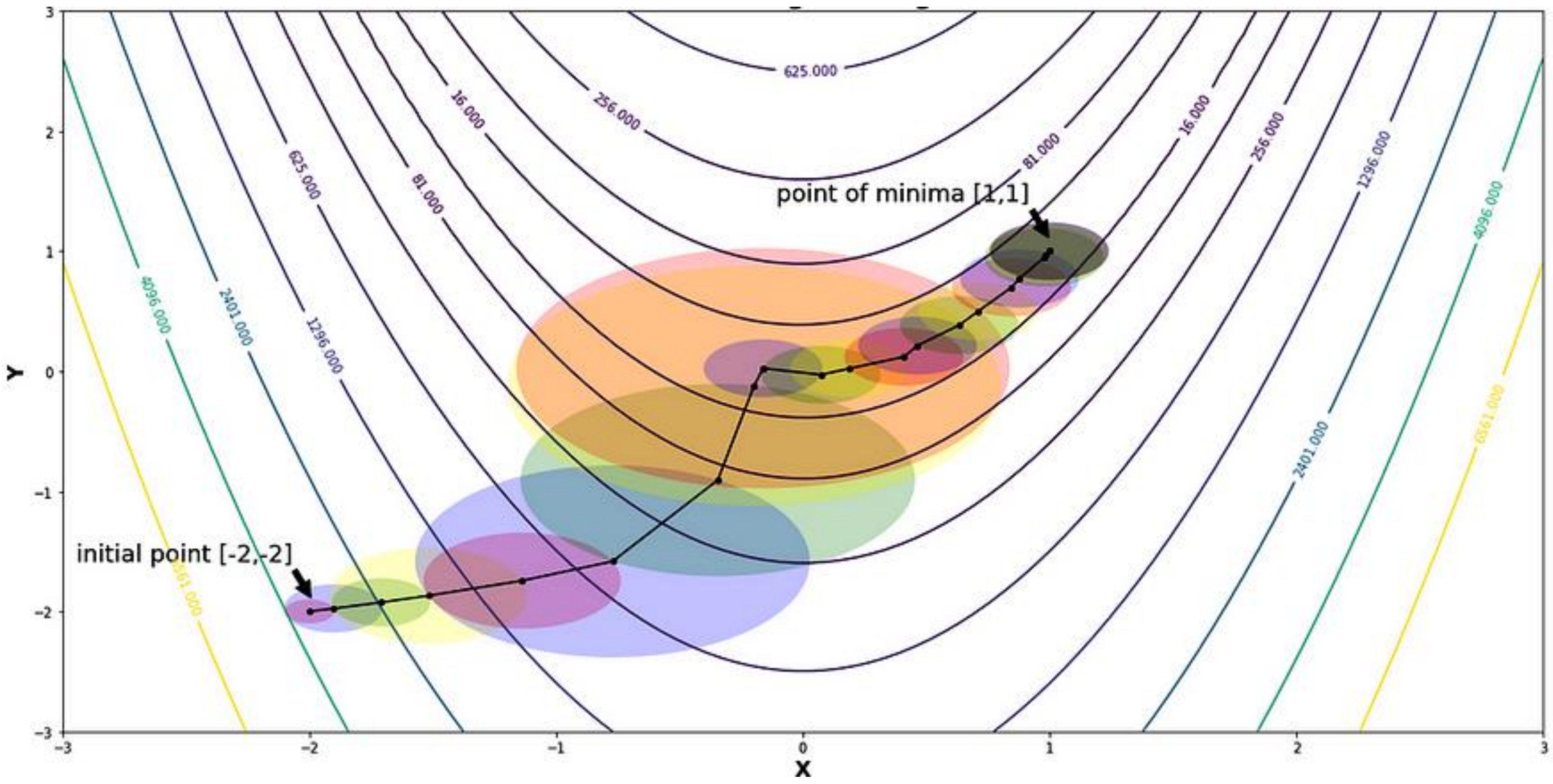
- If $\eta < \eta_1$, then $\delta \leftarrow \delta / \gamma_1$, for $\gamma_1 > 1$
- If $\eta > \eta_2$, then $\delta \leftarrow \delta \cdot \gamma_2$, for $\gamma_2 > 1$

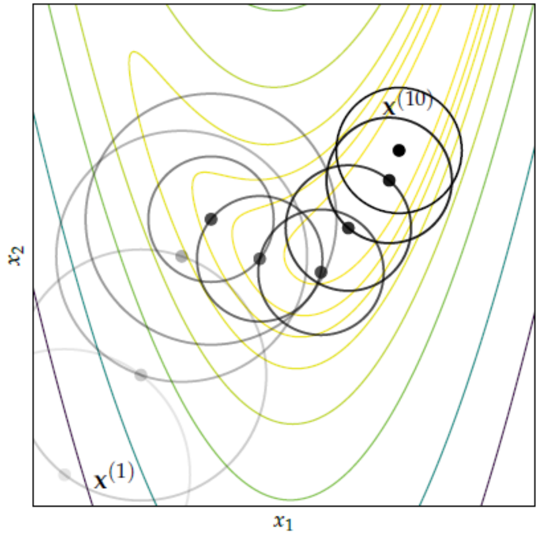
Trust region method

Example

The Rosenbrock function $f(x) = (a - x_1)^2 + b(x_2 - x_1^2)^2$

Global minimum at $\mathbf{x}^* = (a, a^2)$





$a = 1, b = 5$

Trust region subproblem

How to solve?

Fix k and assume $\mathbf{x}_{k-1}^*$ is given. Let us re-write

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}_{k-1}^*) + \nabla f(\mathbf{x}_{k-1}^*)^T \cdot (\mathbf{x} - \mathbf{x}_{k-1}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{k-1}^*)^T \cdot \nabla^2 f(\mathbf{x}_{k-1}^*) \cdot (\mathbf{x} - \mathbf{x}_{k-1}^*)$$

using

$G_k := \nabla f(\mathbf{x}_{k-1}^*)^T$ (gradient),
$B_k := \nabla^2 f(\mathbf{x}_{k-1}^*)$ (Hessian),
$p_k = \mathbf{x} - \mathbf{x}_{k-1}^*$ (step),
$f_k = f(\mathbf{x}_{k-1}^*)$

as

$$m_k(p_k) = \tilde{f}(p_k + \mathbf{x}_{n-1}^*) = f_k + G_k \cdot p_k + \frac{1}{2} p_k^T \cdot B_k \cdot p_k.$$

We need to solve

$m_k(p_k) \to min,$ subject to $\|p_k\| < \delta$

Trust region subproblem

How to solve? (cont.) Cauchy point

$m_k(p_k) \to min,$ subject to $\|p_k\| < \delta$

$$\color{red}{m_k}(p_k) = f_k + G_k \cdot p_k + \frac{1}{2} p_k^T \cdot B_k \cdot p_k$$

Define a **Cauchy point** p_k^C via the following steps:

1. Find the point p_k^ℓ that minimizes the **linear** part of m_k:

$$p_k^\ell = \arg\min_{p \in \mathbb{R}^n} (f_k + G_k \cdot p)\,, \qquad \|p\| < \delta$$

Exercise: Show that $p_k^\ell = -\frac{\delta}{\|G_k\|} G_k$.

The point p_k^ℓ is a 'poor' approximation, so:

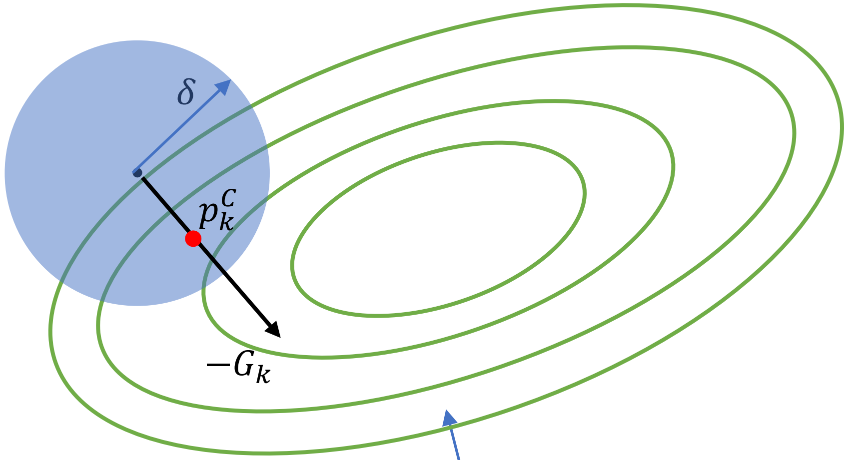
2. Compute the scalar $\tau_k > 0$ that minimizes $m_k\left(\tau_k\, p_k^\ell\right)$ subject to the trust region bound:

$$\tau_k = \arg\min_{\tau \in \mathbb{R}} m_k\left(\tau\, p_k^\ell\right), \qquad \left\|\tau\, p_k^\ell\right\| < \delta$$

Exercise*: Show that $\tau_k = \begin{cases} 1, & \text{if} \quad G_k \cdot B_k \cdot G_k^T \le 0 \\ \min\{1, \hat{\tau}_k\}, & \text{otherwise} \end{cases}$,

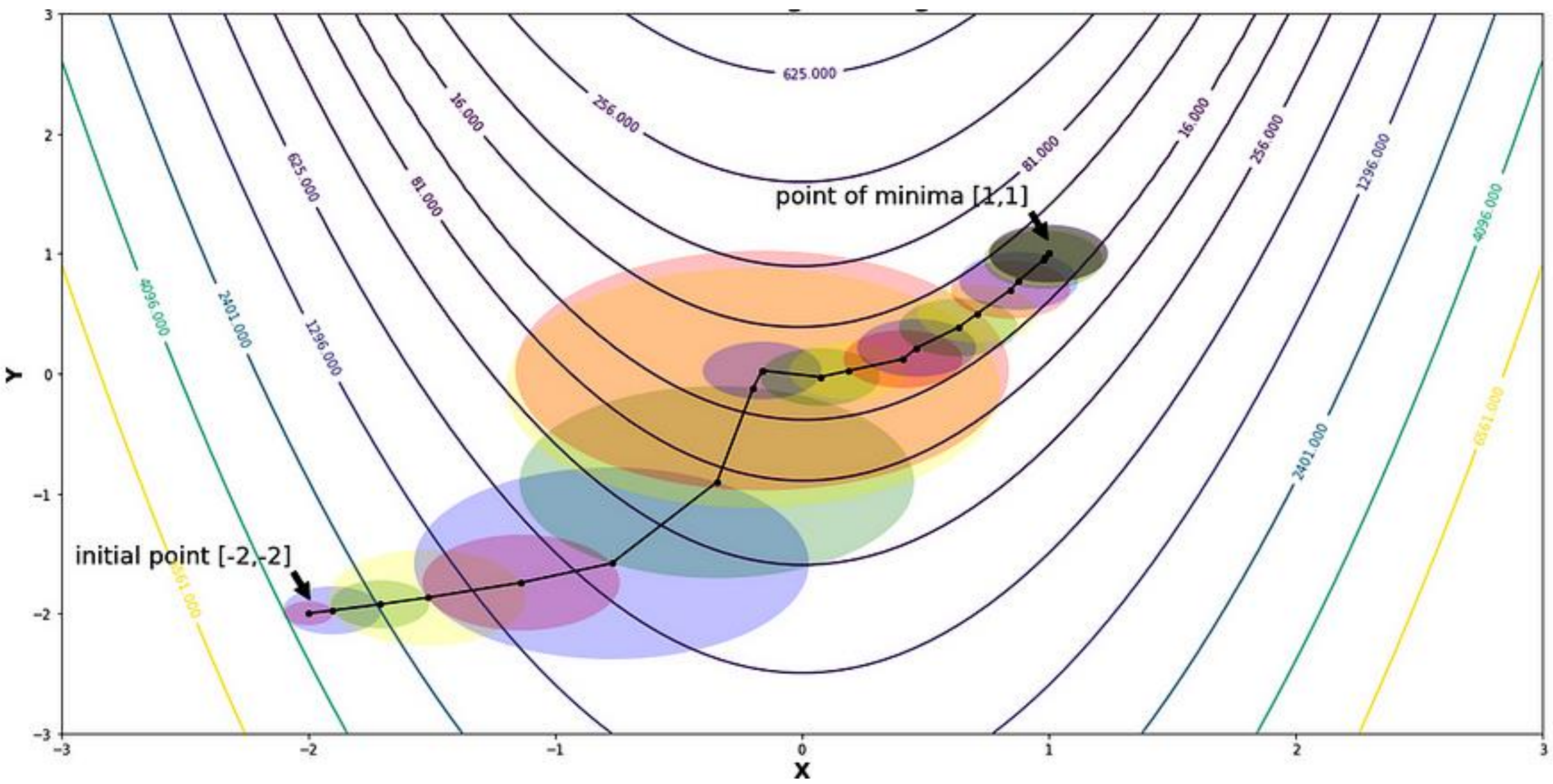
where $\hat{\tau}_k = \frac{\|G_k\|^3}{\delta\ G_k \cdot B_k \cdot G_k^T}$.

3. Set $p_k^C = \tau_k\, p_k^\ell$



Level lines of m_k

Trust region method



Exerciese: Implement 2 steps of Cauchy point search for the Rosenbrock function $f(x_1, x_2) = (1 - x_1)^2 + 5(x_2 - x_1^2)^2$ starting at $(-2, -2)$ and with the trust regions being balls of radius 0.5 for both steps.

Trust region subproblem

How to solve? Cauchy point (cont.)

Summary:

- The Cauchy point is quick to calculate
- It can be shown that the trust region method is globally convergent if its steps $p_n = \mathbf{x}_n^* - \mathbf{x}_{n-1}^*$ attain sufficient reduction in the quadratic approximation
- → The Cauchy point algorithm provides a benchmark against which other methods can be evaluated (such as dog leg method, etc.).

III. Descent direction?

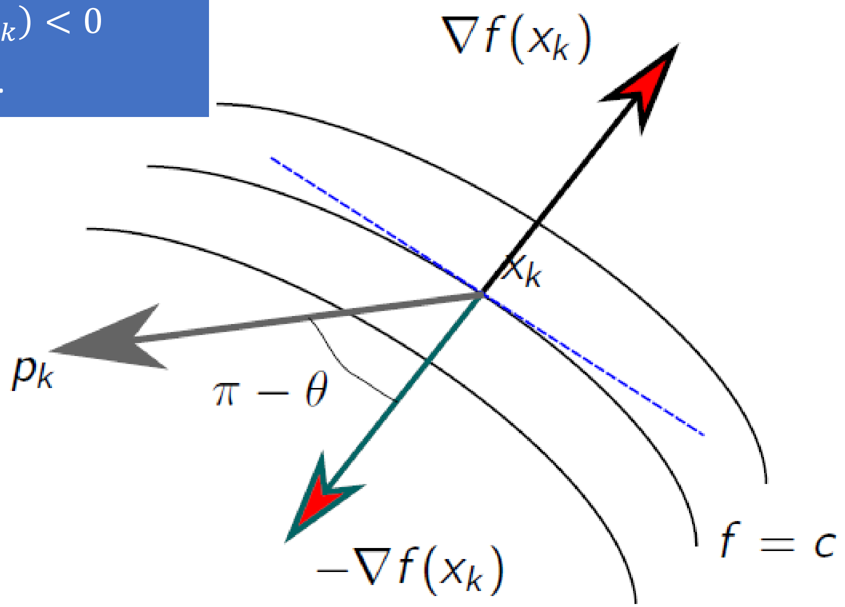
Definition. We say that p_k is a descent direction if $p_k^T \nabla f(\mathbf{x}_k) < 0$.
More generically (in line search methods) we consider

$$p_k = -B_k^{-1}\, \nabla f(\mathbf{x}_k) \qquad \text{with } B_k \text{ positive definite.}$$

Observe that if B_k is positive definite, so is B_k^{-1}.
Therefore, if $p_k = -B_k^{-1} \cdot \nabla f(\mathbf{x}_k)$, then

$$p_k^T \cdot \nabla f(\mathbf{x}_k) = -\left(B_k^{-1} \cdot \nabla f(\mathbf{x}_k)\right)^T \cdot \nabla f(\mathbf{x}_k)$$
$$= -\nabla f(\mathbf{x}_k)^T \cdot \left(B_k^{-1}\right)^T \cdot \nabla f(\mathbf{x}_k) < 0$$

Because $\left(B_k^{-1}\right)^T$ is positive definite.



Notation: $H = \nabla^2 f$

- $B_k = \mathrm{Id}$ (descent method)
- $B_k = Hf(\mathbf{x}_k)$ (Newton method)
- $B_k \approx Hf(\mathbf{x}_k)$ (quasi Newton method)

(Rate of) Convergence?

Steepest descent method = descent along inverse gradient

The ideal case. Assume

$$f(x) = \frac{1}{2}x^T Qx - b^T x$$

where Q is symmetric and positive definite. The gradient is given by $\nabla f(x) = Qx - b$ and so the minimizer $x^\star$ is the (unique) solution of $Qx = b$. Algorithmically,

$$\min_{\alpha \in \mathbb{R}^+} f\left(x - \alpha_k \nabla f\left(x_k\right)\right) \quad \rightarrow \quad \hat{\alpha}_k = \frac{\left(\nabla f\left(x_k\right)\right)^T \nabla f\left(x_k\right)}{\left(\nabla f\left(x_k\right)\right)^T Q \nabla f\left(x_k\right)}$$

where notice that $\nabla f\left(x_k\right) = Qx_k - b$.

Steepest descent method

Definition. Accordingly we have that the steepest decent method with exact line searches writes as

$$x_{k+1} = x_k - \hat{\alpha}_k \nabla f(x_k)$$

To study the rate of convergence we introduce a weighted norm of a vector $x \in \mathbb{R}^n$ as follows

$$||x||_Q^2 = x^T Qx$$

Exercise. If $x^T = (x_1, x_2)$ and $Q = (a_{ij})$ with $i, j = 1, 2$ (symmetric) compute

$$||x||_Q^2.$$

Steepest descent method

Lemma. Assume above notation. We have

$$\frac{1}{2}||x - x^\star||_Q^2 = f(x) - f(x^\star).$$

Proof. The minimizer $x^\star$ satisfies $Qx^\star = b$. Then

$$f(x^\star) = \frac{1}{2}\left((x^\star)^T Qx^\star - 2b^T x^\star\right) = \frac{1}{2}\left((x^\star)^T b - 2b^T x^\star\right) =$$
$$= -\frac{1}{2}b^T x^\star = -\frac{1}{2}(x^\star)^T Qx^\star.$$

where the last equality uses that $Q^T = Q$. Then

$$f(x) - f(x^\star) = \frac{1}{2}\left(x^T Qx - 2b^T x + (x^\star)^T Qx^\star\right) = \frac{1}{2}||x - x^\star||_Q^2$$

since $b^T x = x^\star Qx$.

Steepest descent method

Theorem. When the steepest decent method with exact line searches ($\hat{\alpha}_k$) is applied to the strongly convex quadratic function above then

$$||x_{k+1} - x^\star||_Q^2 \leq \left[\frac{\lambda^n - \lambda_1}{\lambda_n + \lambda_1}\right]^2 ||x_k - x^\star||_Q^2$$

where $0 < \lambda_1 \leq \cdots \lambda_n$ are the eigenvalues of Q.

Remark. The convergence of the steepest decent method under the best conditions, is linear.

Newton's method

Definition. Let f twice differentiable. The **Newton's method** is the line search method defined by

$$p_k = -\left(Hf\left(x_k\right)\right)^{-1} \nabla f\left(x_k\right).$$

Remark. Since $\left(Hf\left(x_k\right)\right)^{-1}$ might not always be positive definite then Newton's method does not always define a descent method. However near the solutions (minimizers) the convergence is quadratic.

Newton's method

Theorem. Assume f is regular (class $\mathcal{C}^3$ is enough) in a neighbourhood of a solution $x^\star$ (minimum of f) where the sufficient optimality conditions hold.
Consider the iteration

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

where p_k is the Newton direction expressed above. Then

(a) $x_k \to x^\star$, if x_0 is close enough to $x^\star$.

(b) The rate of convergence of $\{x_k\}_{k\geq 0}$ is quadratic.

(c) $||\nabla f(x_k)|| \to 0$ quadratically.

Newton's method

Proof: Observe that $\nabla f(x^\star) = 0$ (optimality condition). So,

$$
\begin{aligned}
x_k + p_k - x^\star &= x_k - x^\star - (Hf(x_k))^{-1} \nabla f(x_k) = \\
&= (Hf(x_k))^{-1} \left[Hf(x_k)(x_k - x^\star) - \nabla f(x_k) + \nabla f(x^\star) \right]
\end{aligned}
$$

Observe also that

$$
\begin{aligned}
\nabla f(x^\star) - \nabla f(x_k) &= \int_0^1 \frac{d}{dt} \nabla f(x_k - t(x_k - x^\star)) \; dt = \\
&= \int_0^1 Hf(x_k - t(x_k - x^\star))(x_k - x^\star) \; dt
\end{aligned}
$$

All together implies (L is the Lipschitz constant for $Hf(x)$)

$$
\begin{aligned}
&||Hf(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star))|| \leq \\
&\leq \int_0^1 ||Hf(x_k) - Hf(x_k - t(x_k - x^\star))|| \; ||x_k - x^\star|| \; dt \leq \\
&\leq ||x_k - x^\star||^2 \int_0^1 Lt \; dt = \frac{1}{2} L ||x_k - x^\star||^2
\end{aligned}
$$

Newton's method

Proof (cont.): We go back to

$$||x_k + p_k - x^\star|| = ||\,(Hf\,(x_k))^{-1}\,||\;||\,[Hf\,(x_k)\,(x_k - x^\star) - \nabla f\,(x_k) + \nabla f\,(x^\star)]\,||.$$

We bounded red. Using the regularity of f and th fact that $Hf(x^\star)$ is non singular we have

$$||\,(Hf\,(x_k))^{-1}\,|| \leq 2\,||\,(Hf\,(x^\star))^{-1}\,|| \quad \text{if } ||x_k - x^\star|| < r$$

for some $r > 0$. Finally

$$||x_{k+1} - x^\star|| = ||x_k + p_k - x^\star|| = L||\,(Hf\,(x_k))^{-1}\,||\;||x_k - x^\star||^2 \leq \hat{L}||x_k - x^\star||^2.$$

Choosing x_0 such that $||x_0 - x^\star|| < r$ we can use the inequality inductively to prove (a) and (b). Statement (c) can be proved using similar arguments.

Rate of convergence: general result

Theorem. Suppose f is regular (class $\mathcal{C}^2$ is enough) Consider the iteration $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfying the Wolfe conditions with $c_1 \leq 1$. Assume that the sequence $\{x_k\}_{k \geq 0}$ converges to a point $x^\star$ such that $\nabla f(x^\star) = 0$, $Hf(x^\star)$ is positive definite, and

$$\lim_{k \to \infty} \frac{||\nabla f(x^\star) + Hf(x^\star)(p_k)||}{||p_k||} = 0.$$

Then, the step length $\alpha_k = 1$ is admissible for k large enough and the convergence is linear.