Optimization and Data Science

Lecture 12: Newton Method for Optimization

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Contents

Convergence speed gradient method, quadratic functions

Theorem

For a quadratic function with symmetric positive definite matrix A the gradient method with exact step-size has the R-factor (w.r.t. the Euclidean norm $\|x\|_2 := \sqrt{x^\top x}$):

$$R_{\|\cdot\|_2} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{cond(A) - 1}{cond(A) + 1},$$

where

- $\lambda_{min}, \lambda_{max}$ *are the smallest and biggest eigenvalue of A, respectively,*
- $cond(A) := \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$ *is the* **condition number** *of A.*

Gradient method quadratic function: successive search directions are orthogonal

- Consider again the gradient method for quadratic function. We have

$$d_k = -\nabla f(x_k) = -(Ax_k + b), \quad \text{exact step-size:} \quad \rho_k = \frac{d_k^\top d_k}{d_k^\top A d_k},$$

$$x_{k+1} = x_k + \rho_k d_k$$

$\rightsquigarrow$ next search direction:

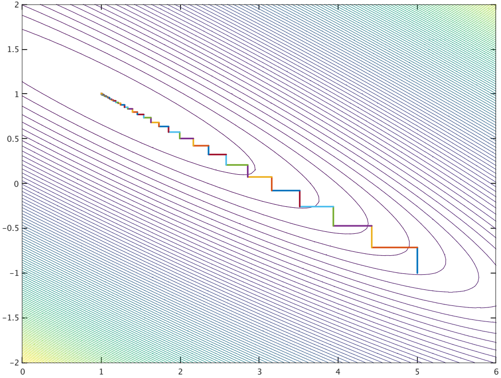
$$d_{k+1} = -(Ax_{k+1} + b) = -\left(A(x_k + \rho_k d_k) + b\right) = -A(x_k + b) - \rho_k A d_k = d_k - \rho_k A d_k$$

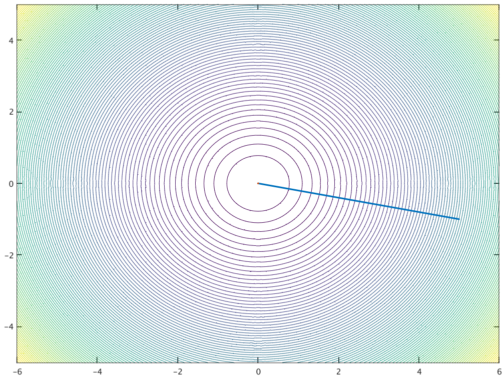
- Now we compute $d_{k+1}^\top d_k$:

$$d_{k+1}^\top d_k = (d_k - \rho_k A d_k)^\top d_k = d_k^\top d_k - \rho_k d_k^\top A d_k = d_k^\top d_k - \frac{d_k^\top d_k}{d_k^\top A d_k} d_k^\top A d_k = 0$$

$\rightsquigarrow$ $d_{k+1}^\top d_k = 0 \Rightarrow d_{k+1} \perp d_k$, two successive search directions are orthogonal.

Gradient method quadratic function: successive search directions are othogonal





- $\lambda_{min} \approx 0.4, \lambda_{max} \approx 17, cond \approx 46, Q \approx 0.96,$ $\qquad \lambda_{min} = \lambda_{max} = 1, cond = 1, Q = 0.$

$\rightsquigarrow$ different curvature of the functions $\rightsquigarrow$ take 2nd derivative into account.

Contents

Newton method: find a root of general nonlinear function $F : \mathbb{R}^n \to \mathbb{R}^n$

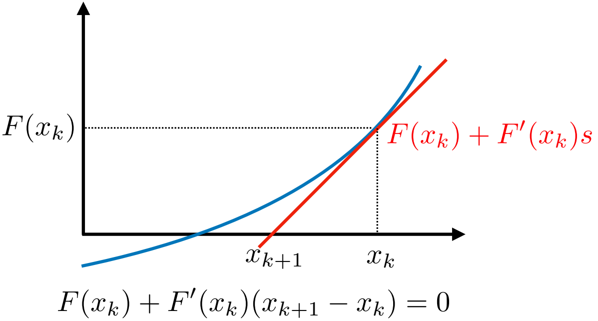
- 1-D: $F : \mathbb{R} \to \mathbb{R}$: Newton method: find zero of tangent at x_k with x-axis

$$F(x_k) + F'(x_k)(\underbrace{x_{k+1} - x_k}_{=:d_k}) = 0$$

⇝ solve (for d_k):

$$F'(x_k)d_k = -F(x_k)$$
$$x_{k+1} = x_k + d_k$$

- Same formula for $F : \mathbb{R}^n \to \mathbb{R}^n$,
- ... but $F'(x_k) \in \mathbb{R}^{n \times n}$ is a matrix now.



Contents

Newton method for optimization

- Newton method for $F : \mathbb{R}^n \to \mathbb{R}^n$: solve (for d_k):

$$F'(x_k)d_k = -F(x_k).$$

- First order necessary condition: $\nabla f(x) = 0$.

$\rightsquigarrow$ consider $F = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$: solve

$$\nabla^2 f(x_k)d_k = -\nabla f(x_k), \tag{1}$$

- ... again with the Hessian matrix:

$$\nabla^2 f(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$$

- The solution d_k of (1) is called **Newton direction**.

A different view on Newton's method

- We consider a general nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$.
- Assume we have (somehow) computed an iterate $x_k \in \mathbb{R}^n$.
- We approximate f in the vicinity of x_k by Taylor expansion

$$f(x_k + d) \approx \underbrace{f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_k) d}_{=:f_k(d)}, \quad d \in \mathbb{R}^n.$$

- f_k is a quadratic function:

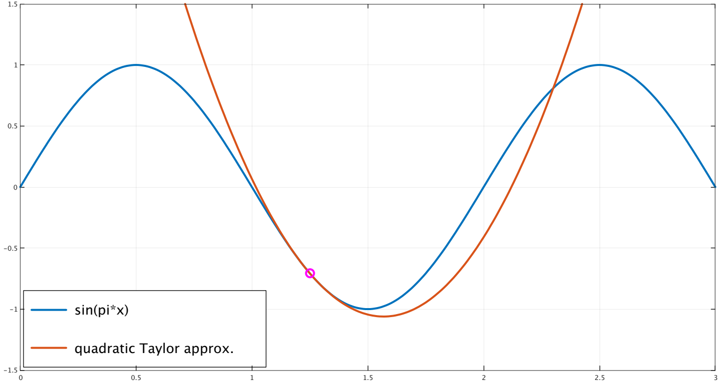
$$f_k(d) = f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_k) d = \frac{1}{2} d^\top A d + b^\top d + c.$$

- The approximation is “good” if d is “small”.

A different view on Newton's method

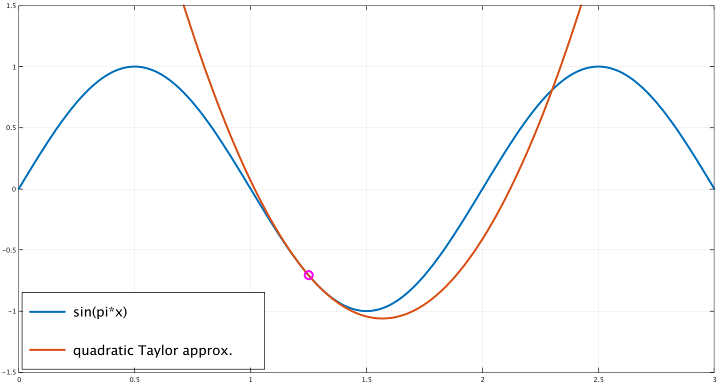
- The quadratic approximation f_k is "good" if d is "small":

$$f_k(d) = f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_k) d = \frac{1}{2} d^\top A d + b^\top d + c.$$



1-D example

$$f(x) = \sin(\pi x), \quad x_k = \frac{5}{4}, \quad f_k(d) = \sin\left(\pi\frac{5}{4}\right) + \pi\cos\left(\pi\frac{5}{4}\right)d - \frac{1}{2}\pi^2\sin\left(\pi\frac{5}{4}\right)d^2$$



A different view on Newton's method

- We approximate f in the vicinity of the current iterate x_k by the quadratic function

$$f(x_k+d) \approx f_k(d) = f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_k) d = \frac{1}{2} d^\top A d + b^\top d + c.$$

- We minimize this approximation w.r.t. d.
- Necessary optimality condition:

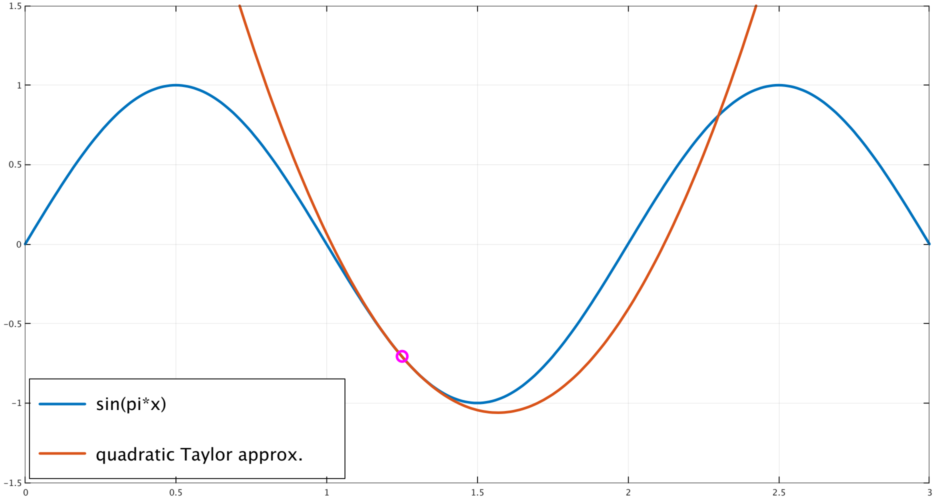
$$\nabla f_k(d) = Ad + b = \nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

- This gives d as solution of

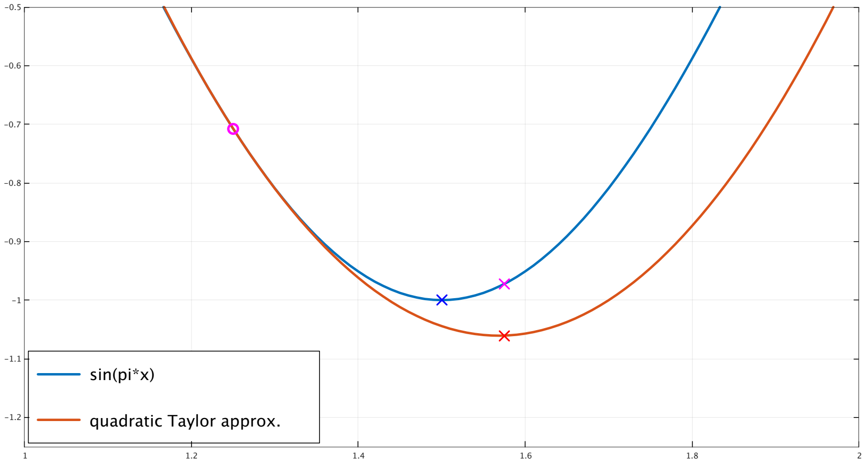
$$\nabla^2 f(x_k) d = -\nabla f(x_k).$$

- $\rightsquigarrow$ d is the Newton direction.
- If $\nabla^2 f(x_k)$ is positive-definite, Newton direction is the unique minimizer of the quadratic approximation of f at the current iterate x_k.

Newton direction: minimizer of quadratic approximation at current iterate

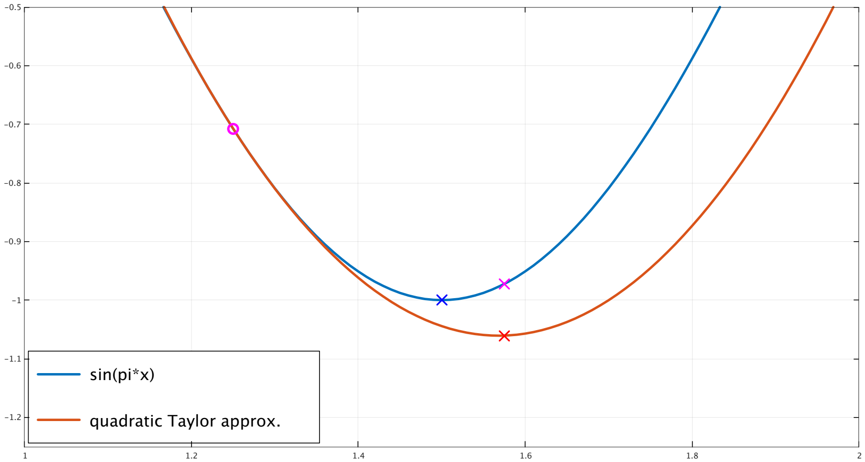


Newton direction: minimizer of quadratic approximation at current iterate



minima of function, quadratic approximation, function value at next full step

Possible benefit of the line-search/globalization in Newton's method



Might be better not to take the full Newton step $\rightsquigarrow$ line search

Contents

Eigenvalues of the Hessian matrix

- If f is twice continuously differentiable, the Hessian matrix $\nabla^2 f(x)$ is symmetric.
- Every symmetric matrix A has only real eigenvalues.
- Every symmetric positive-definite matrix A has only positive eigenvalues:

$$0 < \lambda_{min} \leq \ldots \leq \lambda \leq \ldots \leq \lambda_{max}.$$

- Thus a symmetric positive definite matrix A is invertible (since 0 is no eigenvalue).
- Moreover we can estimate:

$$\|Ay\|_2 \leq \lambda_{max}(A)\|y\|_2,$$
$$\lambda_{min}(A)\|y\|_2^2 \leq y^\top Ay \leq \lambda_{max}(A)\|y\|_2^2 \quad \text{for all } y \in \mathbb{R}^n.$$

- If A is invertible, the eigenvalues of the inverse A^{-1} are the inverse eigenvalues of A:

$$Ax = \lambda x \iff A^{-1}Ax = \lambda A^{-1}x \iff x = \lambda A^{-1}x \iff \frac{1}{\lambda}x = A^{-1}x.$$

Eigenvalues of the inverse Hessian matrix

- We know: The eigenvalues of the inverse A^{-1} are the inverse eigenvalues of A.
- A symmetric positive-definite $\Leftrightarrow$ A^{-1} positive-definite with only positive eigenvalues:

$$0 < \frac{1}{\lambda_{max}(A)} = \lambda_{min}(A^{-1}) \leq \ldots \leq \lambda(A^{-1}) \leq \ldots \leq \lambda_{max}(A^{-1}) = \frac{1}{\lambda_{min}(A)}$$

- ... and:

$$\|A^{-1}y\|_2 \leq \lambda_{max}(A^{-1})\|y\|_2 = \frac{1}{\lambda_{min}(A)}\|y\|_2,$$

$$\frac{1}{\lambda_{max}(A)}\|y\|_2^2 \leq y^\top A^{-1} y \leq \frac{1}{\lambda_{min}(A)}\|y\|_2^2 \quad \text{for all } y \in \mathbb{R}^n.$$

Newton direction for positive-definite Hessian matrix

- We have

$$\|A^{-1}y\|_2 \leq \frac{1}{\lambda_{min}(A)}\|y\|_2,$$

$$\frac{1}{\lambda_{max}(A)}\|y\|_2^2 \leq y^\top A^{-1}y \quad \text{for all } y \in \mathbb{R}^n.$$

- This gives for the check if $d_k = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ is gradient-related:

$$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\|_2\|d_k\|_2} = \frac{\nabla f(x_k)^\top \nabla^2 f(x_k)^{-1}\nabla f(x_k)}{\|\nabla f(x_k)\|_2\|\nabla^2 f(x_k)^{-1}\nabla f(x_k)\|_2}$$

$$\geq \frac{\frac{1}{\lambda_{max}}\|\nabla f(x_k)\|_2^2}{\|\nabla f(x_k)\|_2\frac{1}{\lambda_{min}}\|\nabla f(x_k)\|_2} = \frac{\lambda_{min}(\nabla^2 f(x_k))}{\lambda_{max}(\nabla^2 f(x_k))} =: C_D > 0$$

- $\rightsquigarrow$ Newton directions are gradient-related if Hessian is **uniformly positive-definite**, i.e.

$$0 < c_1 \leq \lambda_{min}(\nabla^2 f(x_k)) \leq \lambda_{max}(\nabla^2 f(x_k)) \leq c_2 < \infty \quad \text{for all } k.$$

Newton method as descent method: Globalized Newton method

- Uniform positive definiteness is a hard condition that we cannot check beforehand.

⇝ We choose $c > 0$ and check in every iteration, if the Newton direction satisfies

$$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\|_2 \|d_k\|_2} \geq c.$$

- If not, we use the negative gradient as search direction instead. It satisfies

$$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\|_2 \|d_k\|_2} = -\frac{-\nabla f(x_k)^\top \nabla f(x_k)}{\|\nabla f(x_k)\|_2 \|\nabla f(x_k)\|_2} = 1.$$

⇝ We have generated a sequence of gradient-related directions with

$$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\| \|d_k\|} \geq c_D = \min\{1, c\}.$$

Newton method as descent method: Globalized Newton method

Algorithm (Globalized Newton method):

1. Fix some parameter $c > 0$.
2. Choose initial guess $x_0 \in \mathbb{R}^n$.
3. For $k = 0, 1, \dots$:
    1. Compute Newton direction d_k, i.e., solve
    $$\nabla^2 f(x_k) d_k = -\nabla f(x_k),$$
    2. If Newton direction is not gradient-related, i.e., if
    $$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\| \|d_k\|} < c,$$
    set $d_k = -\nabla f(x_k)$.
    3. Choose an efficient step-size $\rho_k > 0$.
    4. Set $x_{k+1} = x_k + \rho_k d_k$.

   until a stopping criterion is satisfied.

Convergence result for globalized Newton method

The assumptions of the convergence theorem above are satisfied. But we get more:

Theorem

Let

- *f be twice continuously differentiable,*
- *a subsequence of $(x_k)_{k\in\mathbb{N}}$ converge to x^* where $\nabla f^2(x^*)$ is positive-definite.*

Then

- *x^* is a strict local minimizer,*
- *the whole sequence converges to x^*,*
- *there exists $(q_k)_{k\in\mathbb{N}}, q_k \to 0$, with*

$$\|x_{k+1} - x^*\| \le q_k \|x_k - x^*\| \quad \text{for all } k \in \mathbb{N},$$

i.e., the convergence is Q-superlinear.

Contents

Price to pay: Effort of Newton method

In every Newton step we have to ...

- somehow find a formula for the gradient and the Hessian,
  - either analytically
  - or symbolically using some software
  - or algorithmically (if f is only available as computer program)
- evaluate the gradient:

$$\mathcal{O}(n) \times \text{Effort}(f).$$

- evaluate the Hessian matrix:

$$\mathcal{O}(n^2) \times \text{Effort}(f).$$

- or we find an approximation (if f is only available as black-box, not as source code)
- solve the linear system:

$\mathcal{O}(n^3)$ operations for a dense matrix (less for a sparse matrix)

Contents

How to approximate the gradient?

- Gradient of f at $x \in \mathbb{R}^n$:

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_k}(x) \right)_{k=1}^{n} \in \mathbb{R}^n$$

with components (partial derivatives):

$$\frac{\partial f}{\partial x_k}(x) := \lim_{h \to 0} \frac{f(x + he_k) - f(x)}{h}, \quad k = 1, \dots, n$$

Here $e_k = (0, \dots, 0, \underset{\substack{\uparrow \\ k}}{1}, 0, \dots, 1)$ is the k-the unit vector.

- Finite-difference approximation using a fixed $h > 0$:

$$\frac{\partial f}{\partial x_k}(x) \approx \frac{f(x + he_k) - f(x)}{h}, \quad k = 1, \dots, n.$$

- $\rightsquigarrow$ full gradient approximation takes n additional evaluations of f.

How to approximate the Hessian?

- Hessian is symmetric if f is twice continuously differentiable:

$$\nabla^2 f(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\ldots,n} \in \mathbb{R}^{n\times n}$$

- Finite-difference approximation using a fixed $h > 0$:

$$\begin{aligned}\frac{\partial^2 f}{\partial x_i \partial x_j}(x) &= \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}(x) \approx \frac{\partial}{\partial x_i}\frac{f(x+he_j)-f(x)}{h} \\ &\approx \frac{1}{h}\left(\frac{f(x+he_j+he_i)-f(x+he_i)}{h} - \frac{f(x+he_j)-f(x)}{h} \right) \\ &= \frac{f(x+he_j+he_i)-f(x+he_i)-f(x+he_j)+f(x)}{h^2}, \quad i,j=1,\ldots,n.\end{aligned}$$

- $\rightsquigarrow$ full Hessian approximation takes $\mathcal{O}(n^2)$ additional evaluations of f.

What is important

- Gradient method with exact step-size gives zig-zagging of iterates for quadratic function.

$\rightsquigarrow$ Methods that take into account second derivatives (or approximations) might be useful.

- Newton method for nonlinear equations can be applied on the gradient of the cost function.
- This results in a method that solves a linear system with the Hessian matrix in every step. The solution to this system is called the Newton direction.
- If the Hessian is uniformly positive-definite, the Newton direction is gradient-related.
- The globalized Newton method uses a line search and the Newton direction, if it is gradient-related, and the negative gradient as search direction, if not.
- Under some assumptions, this method shows Q-superlinear convergence.