ROYAL INSTITUTE
OF TECHNOLOGY

Lecture: Nonlinear optimization without constraints

1. Nonlinear optimization without constraints
2. Optimality conditions
3. Optimization algorithms
   - The Gradient method
   - Newton's method
4. Nonlinear least-squares estimation

Local and Global optimas

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \mathbf{R}^n \end{array}$$

Definition 1. $\hat{\mathbf{x}} \in \mathbf{R}^n$ *is a local minimum to the function* f *if there exists a* $\delta > 0$ *such that*

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}),\ \forall \mathbf{x} \in \mathbf{R}^n \text{ such that } \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \delta.$$

Definition 2. $\hat{\mathbf{x}} \in \mathbf{R}^n$ *is a global minimum to the function* f *if*

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}),\ \forall \mathbf{x} \in \mathbf{R}^n$$

First and second order derivatives

The Gradient to f in the point $\mathbf{x}$ is defined as the row-vector

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}(\mathbf{x}) & \dots & \dfrac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

The Hessian to f in the point $\mathbf{x}$ is defined as the symmetric $n \times n$ matrix

$$\mathbf{F}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \dots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \dots & \dfrac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix}$$

Assume from now that f is twice continuously differentiable.

The directional derivative

Consider the function f at the point $\mathbf{x}$ in the direction $\mathbf{d}$, and let $F_{\mathbf{d}}(\alpha) = f(\mathbf{x} + \alpha\mathbf{d})$. It is a function of <u>one</u> variable; the scalar α.

$$F'_{\mathbf{d}}(\alpha) = \lim_{h\to 0} \frac{F_{\mathbf{d}}(\alpha + h) - F_{\mathbf{d}}(\alpha)}{h} = \lim_{h\to 0} \frac{f(\mathbf{x} + \alpha\mathbf{d} + h\mathbf{d}) - f(\mathbf{x} + \alpha\mathbf{d})}{h}$$

$$= \lim_{h\to 0} \frac{f(\mathbf{x} + \alpha\mathbf{d}) + h\nabla f(\mathbf{x} + \alpha\mathbf{d})\mathbf{d} + \frac{1}{2}h^2\mathbf{d}^{\mathsf{T}}\nabla^2 f(\xi)\mathbf{d} - f(\mathbf{x} + \alpha\mathbf{d})}{h}$$

$$= \nabla f(\mathbf{x} + \alpha\mathbf{d})\mathbf{d}$$

Especially it holds that $F'_{\mathbf{d}}(0) = \nabla f(\mathbf{x})\mathbf{d}$ is the directional derivative for f at the point $\mathbf{x}$ and in the direction $\mathbf{d}$.

Descent directions and directional derivatives

Definition 3. $\mathbf{d}$ *is a descent direction to* f *at the point* $\mathbf{x}$ *if there exists an* $\epsilon > 0$ *such that* $f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$ *for all* $t \in (0, \epsilon)$*.*

Descent directions can be characterized using directional derivatives:

Lemma *If* $\nabla f(\mathbf{x})\mathbf{d} < 0$*, then* $\mathbf{d}$ *is a descent direction to* f *at* $\mathbf{x}$*.*

If there are no descent directions to f at the point $\mathbf{x}$ it must hold that $\nabla f(\mathbf{x})\mathbf{d} \geq 0$ for all $\mathbf{d}$, i.e. that $\nabla f(\mathbf{x}) = 0$.

First and second order optimality conditions

Theorem 1 (First order necessary conditions)**.**
If $\hat{\mathbf{x}}$ *is a local minimum to* f *then* $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$*.*

Theorem 2 (Second order necessary conditions)**.**
If $\hat{\mathbf{x}}$ *is a local minimum to* f *then* $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$ *and* $\mathbf{F}(\hat{\mathbf{x}}) \geq \mathbf{0}$
(positive semidefinite),

Theorem 3 (Second order sufficient conditions)**.**
If $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$ *and* $\mathbf{F}(\hat{\mathbf{x}}) > \mathbf{0}$ *(positive definite)*
then $\hat{\mathbf{x}}$ *is a local minimum.*

Example - No descents, but not local minimum

The function $f(x,y) = (y - x^2)(y - 2x^2)$ is zero at $(x^*, y^*) = (0,0)$. It has no descent directions there; if $d = (\alpha, \beta)$ then

$$f(x^* + t\alpha, y^* + t\beta) - f(x^*, y^*) = t^2(\beta^2 - 3t\alpha^2\beta + t^2\alpha^4)$$

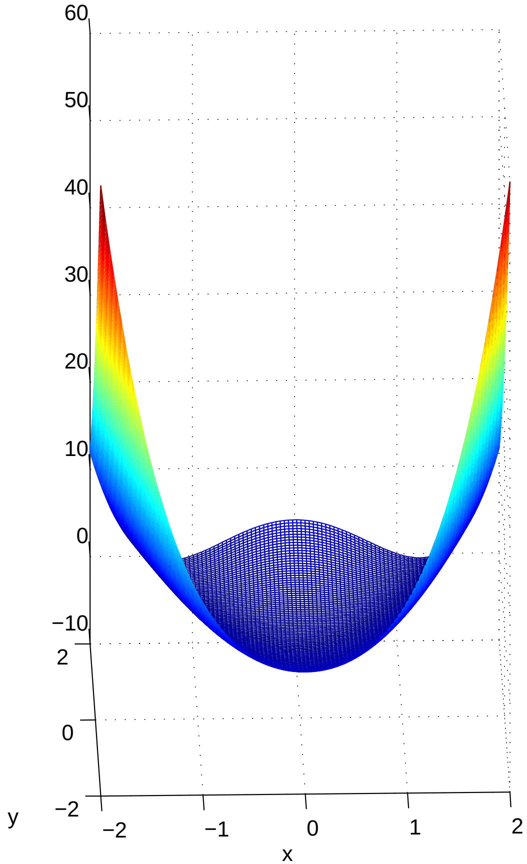
$$= \begin{cases} > 0 & \text{if } t < |\beta|/(3\alpha^2), \beta \neq 0, \alpha \neq 0 \\ 2t^4\alpha^4 > 0 & \text{if } \beta = 0, \\ t^2\beta^2 > 0 & \text{if } \alpha = 0. \end{cases}$$

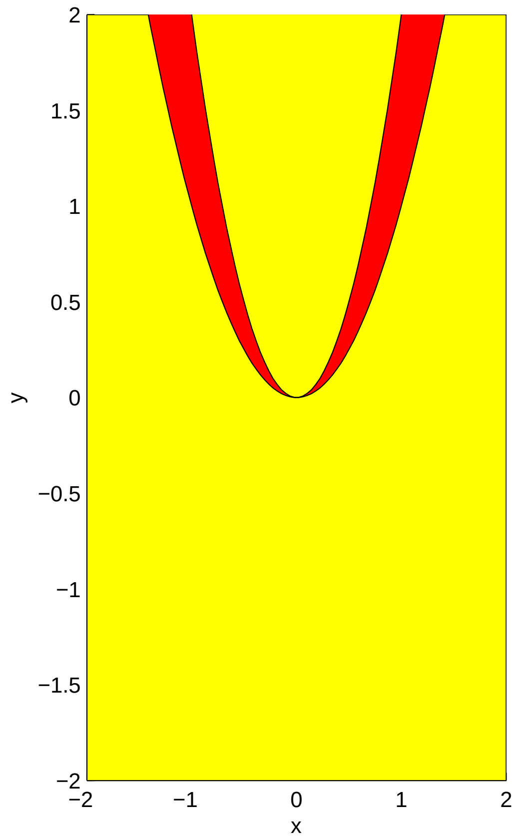
Along no straight line through the origin there is an initial descent.

Note: $f(t, \frac{3}{2}t^2) = -\frac{t^2}{4} < 0$ so (x^*, y^*) is not a local minimum.

Example - Graphical illustration

The function $z = f(x, y)$ is depicted below, in $\mathbf{R}^3$ (left) and in $\mathbf{R}^2$ (right). On the right the function is negative in the red region and positive in the yellow region.





Example - Descents, but no negative directional derivatives

The function $f(x,y) = -(x^4 + y^4)$ is zero at $(x^*, y^*) = (0,0)$.

Note that $\nabla f(x^*, y^*) = \begin{bmatrix} 0 & 0 \end{bmatrix}$, and $F(x^*, y^*) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

So the point (x^*, y^*) satisfies the first and second order necessary conditions for optimality, but not the second order sufficient conditions.

For the direction $d = (\alpha, \beta)^T$

$$f(x^* + t\alpha, y^* + t\beta) - f(x^*, y^*) = -t^4(\alpha^4 + \beta^4) < 0$$

so along all straight lines through the origin there is an initial descent, but no directional derivative $\nabla f(x^*, y^*)d$ is negative.

Note: (x^*, y^*) is in fact the global maximum for f.

Optimization algorithms

We consider two iterative methods for minimization of multivariable functions.

1. The Gradient method (steepest descent)
   - The search direction is determined from the gradient, *i.e.*, first order information.
2. Newton's method.
   - The search direction is determined from the gradient and Hessian, *i.e.*, second order information.

The Gradient method

Idéa: Search in the direction that the function decreases the most $\Rightarrow$ the search direction is determined by $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^\mathsf{T}$.

Algorithm:

(0) Determine starting point $\mathbf{x}^{(0)}$ and let $k = 0$.

(i) Let $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^\mathsf{T}$.

(ii) Check the stopping criterion: If $|\nabla f(\mathbf{x}^{(k)})| \leq \epsilon$ the search is terminated.

(iii) Perform the line search

$$t^{(k)} = \arg\min_{t \geq 0} f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})$$

and let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\mathbf{d}^{(k)}$

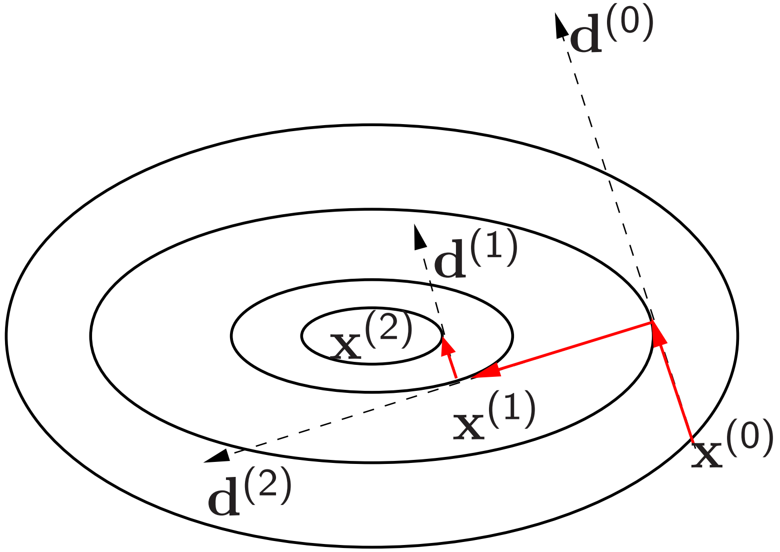
(iv) Update $k = k + 1$ and go to step (i).

Comments on the gradient method

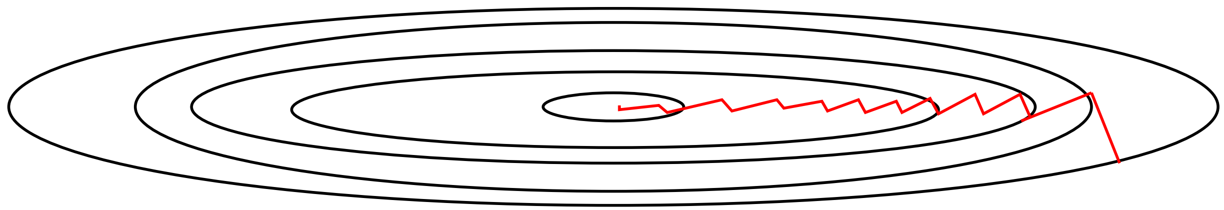
If exact line search is performed, then $\mathbf{d}^{(k+1)} \perp \mathbf{d}^{(k)}$.

Proof: If $\varphi(t) = f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})$, then

$$0 = \varphi'(t^{(k)}) = \nabla f(\mathbf{x}^{(k)} + t^{(k)}\mathbf{d}^{(k)})^{\mathsf{T}}\mathbf{d}^{(k)} = -(\mathbf{d}^{(k+1)})^{\mathsf{T}}\mathbf{d}^{(k)}$$



Orthogonal search directions



This can lead to slow convergence

Example

Let $f(\mathbf{x}) = x_1^2 + 2x_2^2 + x_1x_2 + x_2$ and $\mathbf{x}^{(0)} = (0, 0)$.

Then $\nabla f(\mathbf{x}) = (2x_1 + x_2, 4x_2 + x_1 + 1)$, $\mathbf{d}^{(0)} = -\nabla f(\mathbf{x}^{(0)}) = (0, -1)$.

Perform exact line search:

$$\varphi_0(t) = f(\mathbf{x}^{(0)} + t\mathbf{d}^{(0)}) = f(0, -t) = 2t^2 - t,$$

minimized for $t = 1/4$, giving $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + 1/4\mathbf{d}^{(0)} = (0, -1/4)$.

The next search direction is then $\mathbf{d}^{(1)} = -\nabla f(\mathbf{x}^{(1)}) = (1/4, 0)$.

Perform exact line search:

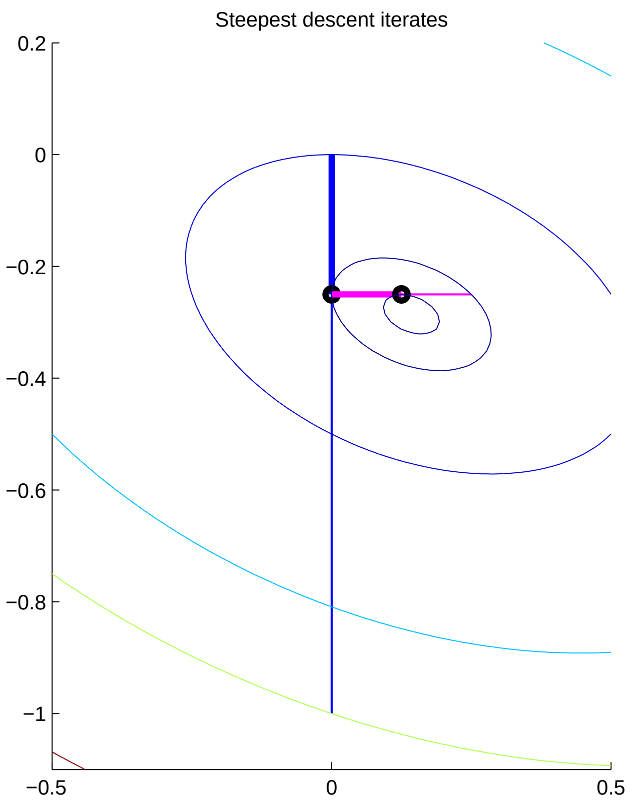
$$\varphi_1(t) = f(\mathbf{x}^{(1)} + t\mathbf{d}^{(1)}) = f(t/4, -1/4) = t^2/16 - t/16 - 1/8,$$

minimized for $t = 1/2$, giving $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + 1/2\mathbf{d}^{(1)} = (1/8, -1/4)$.

Example - Graphical illustration

For every iteration we approach the minimum which is located at $x = -H^{-1}c = (1/7, -2/7)$

$$\mathbf{x}^{(0)} = (0,0), \quad \mathbf{x}^{(1)} = (0,-1/4), \quad \mathbf{x}^{(2)} = (1/8,-1/4).$$



Line search

We let $\varphi(t) = f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})$. The line search corresponds to solving

$$\min_{t \geq 0} \varphi(t).$$

The line search is usually performed approximatively. We present two methods:

1. The bisection method
2. Newton's method

The bisection method

The bisection method uses first order information to search for a point where $\varphi'(t^{(k)}) \approx 0$.

Algorithm:

(0) Let $\alpha_0 = 0$ and $\beta_0 = t_{\max}$, where $t_{\max}$ is an upper limit such that $\varphi'(t_{\max}) > 0$.

(i) $t_k = \dfrac{\alpha_k + \beta_k}{2}$.

(ii) If $|\varphi'(t_k)| \leq \epsilon$ then $t^{(k)} = t_k$. Finished!

(iii) If $\varphi'(t_k) < 0$ then $\alpha_{k+1} = t_k$ and $\beta_{k+1} = \beta_k$.
If $\varphi'(t_k) \geq 0$ then $\alpha_{k+1} = \alpha_k$ and $\beta_{k+1} = t_k$.

(iv) $k = k + 1$. Go to (i).

Newton's method (for line search)

Newton's method uses first and second order information to search for a point where $\varphi'(t^{(k)}) \approx 0$.

Let $t_0 = 0$ and perform the iteration

$$t_{k+1} = t_k - \frac{\varphi'(t_k)}{\varphi''(t_k)}$$

until $|\varphi'(t_k)| \leq \epsilon$. The optimal point is approximatively $t^{(k)} = t_k$.

This method is described in more generality next.

Newton's method

The idéa behind Newton's method is to approximate $f(\mathbf{x})$ with a second order Taylor expansion.

Let $\mathbf{x} = \mathbf{x}^{(k)} + \mathbf{d}$. Newton's method uses the approximation

$$\min_{\mathbf{x}} f(\mathbf{x}) \approx \min_{\mathbf{d}} f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})\mathbf{d} + \frac{1}{2}\mathbf{d}^{\mathsf{T}}\mathbf{F}(\mathbf{x}^{(k)})\mathbf{d}$$

If $\mathbf{F}(\mathbf{x}^{(k)}) > 0$ (positive definite) the minimum $\mathbf{d}^{(k)}$ satisfies

$$\mathbf{F}(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^{\mathsf{T}}$$

If it is not positive definite, then let $\mathbf{H}(\mathbf{x}^{(k)}) = \mathbf{F}(\mathbf{x}^{(k)}) + \mu I$, where $\mu > 0$ is large enough such that $\mathbf{H}(\mathbf{x}^{(k)}) > 0$ and then use the search direction $\mathbf{d}^{(k)}$ satisfying

$$\mathbf{H}(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^{\mathsf{T}}$$

Newton's algorithm:

(0) Determine starting point $\mathbf{x}^{(0)}$ and let $k = 0$.

(i) Check the stopping criterion: $\|\nabla f(\mathbf{x}^{(k)})\| \leq \epsilon$

(ii) Determine search direction

$$\mathbf{F}(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^{\mathsf{T}}$$

(iii) Let $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\mathbf{d}^{(k)}$, where $t^{(k)}$ is the largest of the numbers $1,\ 1/2,\ 1/4, \ldots$ such that $f(\mathbf{x}^{(k)} + t^{(k)}\mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)})$

(iv) $k = k + 1$. Go to (i).

Comment 1. (iii) *can be replaced with a line search. This is especially recommended if* $f(\cdot)$ *is not a convex function.*

Quadratic convergence of Newton's method

(Not in the course curriculum, but you should know about it)

Let $f : S \to \mathbf{R}$, where $S \subset \mathbf{R}^n$ is open and convex. Assume that $\nabla^2 f$ is Lipschitz continuous on S, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \forall x, y \in S, \text{for some} L < \infty.$$

Let x_* be a minimizer of f in S and assume that $\nabla^2 f(x_*)$ is positive definite.

If $\|x_0 - x_*\|$ is sufficently small, then $\{x^{(k)}\}$ defined by $x^{(k+1)} = x^{(k)} + d^{(k)}$ converges quadratically to x_*, *i.e.*,

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x_*\|}{\|x^{(k)} - x_*\|^2} = C < \infty.$$

Example

Let $f(x) = \sqrt{1+x^2}$ and $\mathbf{x}^{(0)} = 2$.

Then $\nabla f(\mathbf{x}) = \frac{x}{\sqrt{1+x^2}}$, $F(\mathbf{x}) = \frac{1}{(1+x^2)^{3/2}}$.

Since the Hessian is positive definite for all x, f is convex.

First iteration

Let $d^{(0)} = -\left(\nabla^2 f(\mathbf{x}^{(0)})\right)^{-1} \nabla f(\mathbf{x}^{(0)}) = -5\sqrt{5} \cdot 2/\sqrt{5} = -10$.

Try first with unit step, which gives function value

$$f(\mathbf{x}^{(0)} + \mathbf{d}^{(0)}) = f(2 + (-10)) = \sqrt{1 + (-8)^2} = \sqrt{65},$$

At the starting point we had $f(\mathbf{x}^{(0)}) = \sqrt{5}$, which was much better.

We have to reduce the steplength. Since $d^{(0)}$ is a descent direction the function should decrease for small enough steps.

Reduce the step length by 1/2:

$$f(\mathbf{x}^{(0)} + \frac{1}{2}\mathbf{d}^{(0)}) = f(2 + (-5)) = \sqrt{1 + (-3)^2} = \sqrt{10},$$

Reduce the step length by 1/4:

$$f(\mathbf{x}^{(0)} + \frac{1}{4}\mathbf{d}^{(0)}) = f(2 + (-2.5)) = \sqrt{1 + (-1/2)^2} = \sqrt{5}/2,$$

which is an improvement. Let $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + 1/4\mathbf{d}^{(0)} = -1/2$.

Second iteration

Then $d^{(1)} = -\left(\nabla^2 f(\mathbf{x}^{(1)})\right)^{-1} \nabla f(\mathbf{x}^{(1)}) = -\frac{5\sqrt{5}}{8}\frac{-1}{\sqrt{5}} = 5/8$.

Try first unit step, which gives function value

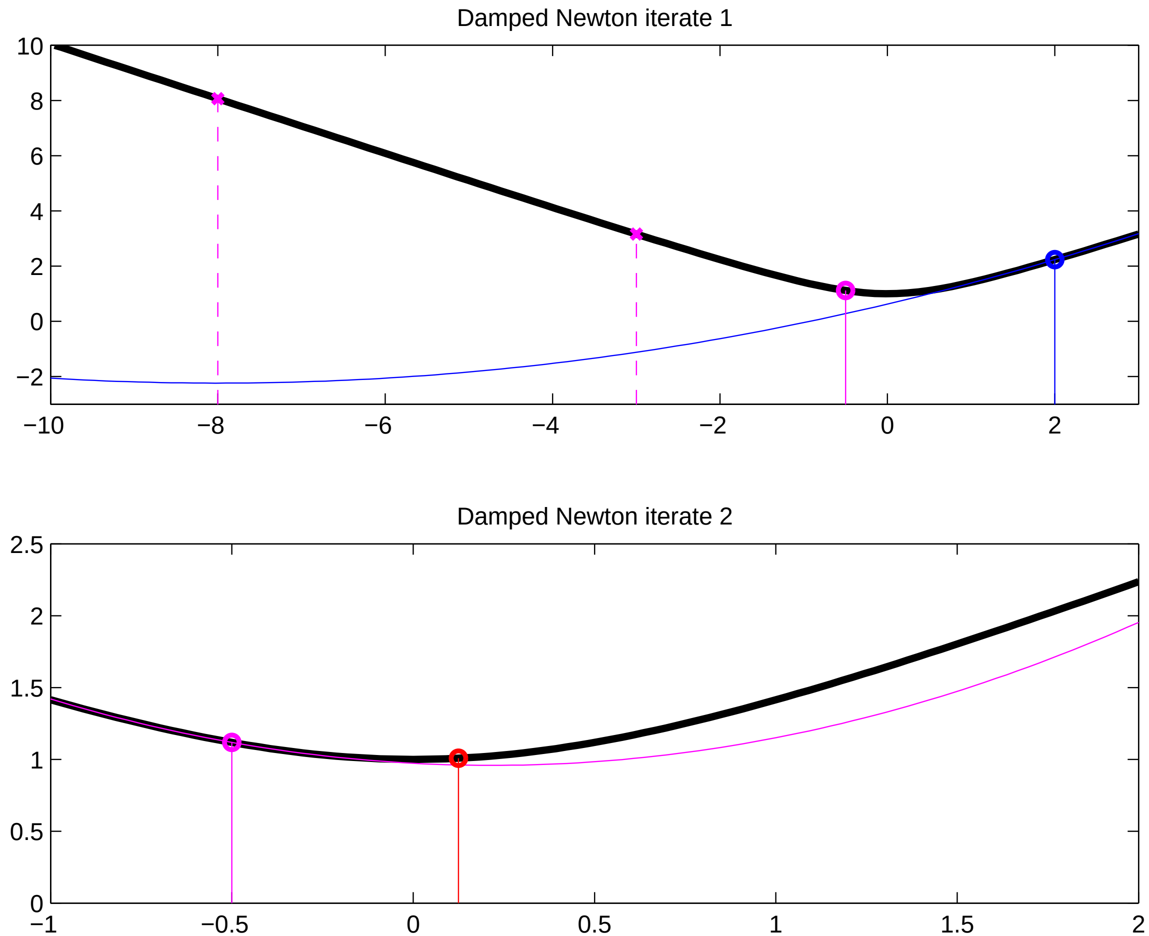
$$f(\mathbf{x}^{(1)} + \mathbf{d}^{(1)}) = f(-1/2 + (5/8)) = \sqrt{1 + (1/8)^2} = \sqrt{65}/\sqrt{64},$$

which is better than $f(\mathbf{x}^{(1)}) = \sqrt{5}/2$. Let $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{d}^{(1)} = 1/8$.

Example - Graphical illustration

For every iteration we approach the minimum which is located at $x = 0$.

$$\mathbf{x}^{(0)} = 0, \quad \mathbf{x}^{(1)} = -1/2, \quad \mathbf{x}^{(2)} = 1/8.$$



Nonlinear least-squares estimation

Problem Find $\mathbf{x}$ so that (approximatively)

$$\begin{aligned} h_1(\mathbf{x}) &= 0 \\ h_2(\mathbf{x}) &= 0 \\ &\vdots \\ h_m(\mathbf{x}) &= 0 \end{aligned}$$

Idéa: Solve the nonlinear least-squares problem:

$$\min f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m} h_i(\mathbf{x})^2 = \frac{1}{2}\mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}) \quad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{bmatrix} \tag{1}$$

If $f(\hat{\mathbf{x}}) \approx 0$ it holds that $h_i(\hat{\mathbf{x}}) \approx 0$, $i = 1, \ldots, m$.

Gauss-Newton's method

We consider two derivations of the Gauss-Newton's method

(for solving (1).)

Method 1: If we use that

$$\mathbf{h}(\mathbf{x}^{(k)} + \mathbf{d}) \approx \mathbf{h}(\mathbf{x}^{(k)}) + \nabla \mathbf{h}(\mathbf{x}^{(k)})\mathbf{d}$$

we get the approximation

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{h}(\mathbf{x})^{\mathsf{T}}\mathbf{h}(\mathbf{x}) \approx \min_{\mathbf{d}} \frac{1}{2}|\nabla h(\mathbf{x}^{(k)})\mathbf{d} + \mathbf{h}(\mathbf{x}^{(k)})|^2$$

This is a least-squares problem in standard form, whos solution is given by the normal equations:

$$\nabla \mathbf{h}(\mathbf{x}^{(k)})^{\mathsf{T}} \nabla \mathbf{h}(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = -\nabla \mathbf{h}(\mathbf{x}^{(k)})^{\mathsf{T}}\mathbf{h}(\mathbf{x}^{(k)})$$

The next iteration point is then given by $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\mathbf{d}^{(k)}$ where $t^{(k)}$ is for example determined with a line search.

Method 2: Use the Newton direction. With $f(\mathbf{x}) = \frac{1}{2}\mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x})$ we get

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{m} h_i(\mathbf{x})\nabla h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\mathsf{T}\nabla\mathbf{h}(\mathbf{x})$$

$$\begin{aligned}\mathbf{F}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) &= \sum_{i=1}^{m}(\nabla h_i(\mathbf{x})^\mathsf{T}\nabla h_i(\mathbf{x}) + h_i(\mathbf{x})\nabla^2 h_i(\mathbf{x})) \\ &= \nabla\mathbf{h}(\mathbf{x})^\mathsf{T}\nabla\mathbf{h}(\mathbf{x}) + \sum_{i=1}^{m} h_i(\mathbf{x})\nabla^2 h_i(\mathbf{x}))\end{aligned}$$

The Newton direction is given by

$$\left(\nabla\mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T}\nabla\mathbf{h}(\mathbf{x}^{(k)}) + \sum_{i=1}^{m} h_i(\mathbf{x}^{(k)})\nabla^2 h_i(\mathbf{x}^{(k)})\right)\mathbf{d}^{(k)} = -\mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T}\nabla\mathbf{h}(\mathbf{x}^{(k)})$$

If we do the approximation $h_i(\mathbf{x}^{(k)}) \approx 0$ we get

$$\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)}) \mathbf{d}^{(k)} = -\mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)})$$

which coincides with **Method 1**.

The next iteration point is given by $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \mathbf{d}^{(k)}$ where $t^{(k)}$, for example, is determined by a line search.

Reading instructions

- Chapter 12-17 in the book.