Optimization
– Smooth optimization –

Nelly Pustelnik

CNRS, Laboratoire de Physique de l'ENS de Lyon, France

(several slides in this part are written in collaboration with **Elisa Riccietti** from LIP, ENS de Lyon)

Solution of a minimization problem

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to \left]-\infty, +\infty\right]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a **local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that
$$(\forall x \in O \cap C) \qquad f(\widehat{x}) \leq f(x).$$
- $\widehat{x}$ is a **(global) minimizer** of f if
$$(\forall x \in C) \qquad f(\widehat{x}) \leq f(x).$$

Solution of a minimization problem

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to \left]-\infty, +\infty\right]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x}$ is a **strict local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that

$$(\forall x \in (O \cap C) \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$

- $\widehat{x}$ is a **strict (global) minimizer** of f if

$$(\forall x \in C \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$

Differentiable functions

Let $f: \mathbb{R}^N \to \left]-\infty, +\infty\right]$ be a proper differentiable function in the neighborhood of $x \in \mathbb{R}^N$.

The **directional derivative** of f at x with respect to the direction $y \in \mathbb{R}^N$ is defined as:

$$\langle \nabla f(x) \mid y \rangle = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}.$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)

Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$. $\widehat{x}$ is a global minimizer of f i.e

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}}\, f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)
Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$. $\widehat{x}$ is a global minimizer of f i.e

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}}\, f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

Proof ($\Rightarrow$): Let $\epsilon \in \mathbb{R}^N$. We set, for every $\alpha \in \mathbb{R}$, $g(\alpha) = f(\widehat{x} + \alpha\epsilon)$. Then

$$\frac{dg(\alpha)}{d\alpha} = \langle \epsilon, \nabla f(\widehat{x} + \alpha\epsilon) \rangle$$

Leading to

$$\begin{aligned} \frac{dg(0)}{d\alpha} &= \langle \epsilon, \nabla f(\widehat{x}) \rangle \\ &= \lim_{\alpha \searrow 0} \frac{f(\widehat{x} + \alpha\epsilon) - f(\widehat{x})}{\alpha} \geq 0 \quad \textbf{(because } \widehat{x} \textbf{ is a minimizer of } f\textbf{)} \end{aligned}$$

Thus

$$\boxed{\langle \epsilon, \nabla f(\widehat{x}) \rangle \geq 0}$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)

Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$. $\widehat{x}$ is a global minimizer of f i.e

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}}\, f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

<u>Proof</u> ($\Rightarrow$): Let $\epsilon \in \mathbb{R}^N$. We set, for every $\alpha \in \mathbb{R}$, $g(\alpha) = f(\widehat{x} - \alpha\epsilon)$. Then

$$\frac{dg(\alpha)}{d\alpha} = \langle -\epsilon, \nabla f(\widehat{x} - \alpha\epsilon)\rangle$$

Leading to

$$\begin{aligned}\frac{dg(0)}{d\alpha} &= \langle -\epsilon, \nabla f(\widehat{x})\rangle \\ &= \lim_{\alpha \searrow 0} \frac{f(\widehat{x} - \alpha\epsilon) - f(\widehat{x})}{\alpha} \geq 0 \quad \text{(because } \widehat{x} \text{ is a minimizer of } f\text{)}\end{aligned}$$

Thus

$$\boxed{\langle \epsilon, \nabla f(\widehat{x})\rangle \leq 0}$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)

Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$. $\widehat{x}$ is a global minimizer of f i.e

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

<u>Proof</u> $(\Leftarrow)$: f being a convex function, this yields to

$$\begin{aligned}
(\forall (x,z) \in \mathbb{R}^N \times \mathbb{R}^N)(\forall \alpha \in [0,1]) \quad & f((1-\alpha)x + \alpha z) \leq (1-\alpha) f(x) + \alpha f(z) \\
\Leftrightarrow \quad & f(x + \alpha(z-x)) \leq (1-\alpha) f(x) + \alpha f(z) \\
\Leftrightarrow \quad & \frac{f(x+\alpha(z-x)) - f(x)}{\alpha} \leq f(z) - f(x)
\end{aligned}$$

Thus

$$\lim_{\alpha \searrow 0} \frac{f(x+\alpha(z-x)) - f(x)}{\alpha} = \langle z - x, \nabla f(x) \rangle \leq f(z) - f(x)$$

If $\nabla f(x) = 0$, then

$$(\forall z \in \mathbb{R}^N) \quad f(z) \geq f(\widehat{x})$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)
Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$. $\widehat{x}$ is a global minimizer of f i.e

$$\widehat{x} \in \underset{x\in\mathbb{R}^N}{\operatorname{Argmin}}\, f(x) \qquad \Leftrightarrow \qquad \nabla f(\widehat{x}) = 0$$

<u>Proof</u> $(\Leftarrow)$: f being a convex function, this yields to

$$\begin{aligned}(\forall (x,z) \in \mathbb{R}^N \times \mathbb{R}^N)(\forall \alpha \in [0,1]) \quad & f((1-\alpha)x + \alpha z) \leq (1-\alpha) f(x) + \alpha f(z)\\ \Leftrightarrow \quad & f(x + \alpha(z-x)) \leq (1-\alpha) f(x) + \alpha f(z)\\ \Leftrightarrow \quad & \frac{f(x+\alpha(z-x)) - f(x)}{\alpha} \leq f(z) - f(x)\end{aligned}$$

Thus

$$\lim_{\alpha \searrow 0} \frac{f(x+\alpha(z-x)) - f(x)}{\alpha} = \boxed{\langle z - x, \nabla f(x)\rangle \leq f(z) - f(x)}$$

$\rightarrow$ **caracterization of the convexity**.

Iterative methods

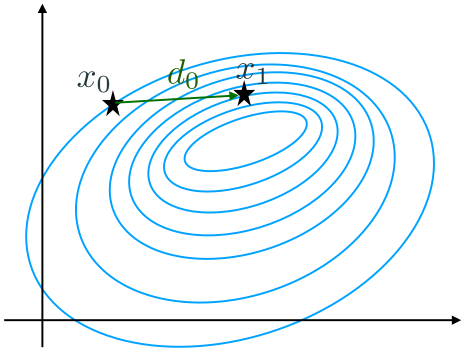
- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.

- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

  where

  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.



Iterative methods

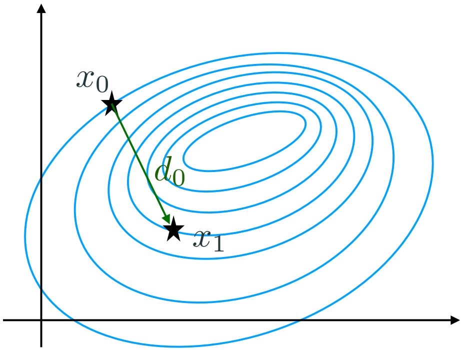
- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.

- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

where

  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.



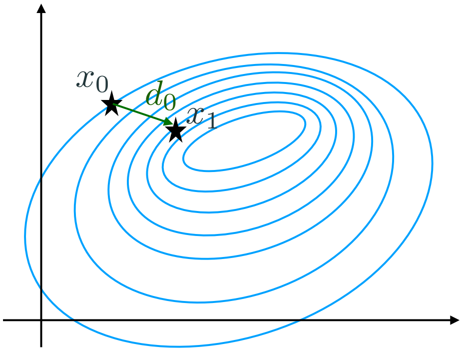
Iterative methods

- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.
- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

where

- $t_k > 0$: step-length,
- $d_k \in \mathbb{R}^N$: step direction.



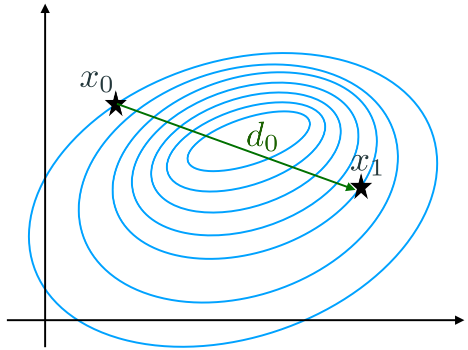
Iterative methods

- Goal: build a sequence $(x_k)_{k \in \mathbb{N}}$ that converges to $\widehat{x}$.
- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

where

  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.



Iterative methods

- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.

- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

  where

  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.

- The choice of d_k is such that it is possible to find a $t_k > 0$ satisfying

$$f(x_k + t_k d_k) = \boxed{f(x_{k+1}) < f(x_k)}$$

Steepest descent

Taylor expansion: Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable, then for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|)$$

Steepest descent

Taylor expansion: Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable, then for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|)$$

- Considering iterations of the form:

$$x_{k+1} = x_k + t_k d_k$$

we get

$$f(x_k + t_k d_k) = f(x_k) + \langle \nabla f(x_k), t_k d_k \rangle + o(\|t_k d_k\|)$$

Steepest descent

Taylor expansion: Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable, then for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|)$$

- Considering iterations of the form:

$$x_{k+1} = x_k + t_k d_k$$

  we get

$$f(x_k + t_k d_k) = f(x_k) + \langle \nabla f(x_k), t_k d_k \rangle + o(\|t_k d_k\|)$$

- In order to have $f(x_{k+1}) < f(x_k)$, we need

$$\langle \nabla f(x_k), d_k \rangle < 0$$

Steepest descent

Taylor expansion: Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable, then for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|)$$

- Considering iterations of the form:

$$x_{k+1} = x_k + t_k d_k$$

we get

$$f(x_k + t_k d_k) = f(x_k) + \langle \nabla f(x_k), t_k d_k \rangle + o(\|t_k d_k\|)$$

- In order to have $f(x_{k+1}) < f(x_k)$, we need

$$\langle \nabla f(x_k), d_k \rangle < 0$$

- The most natural choice is

$$d_k = -\nabla f(x_k)$$

Steepest descent

Taylor expansion: Let $f : \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable, then for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + o(\|\mathbf{y} - \mathbf{x}\|)$$

- Considering iterations of the form:

$$x_{k+1} = x_k + t_k d_k$$

  we get

$$f(x_k + t_k d_k) = f(x_k) + \langle \nabla f(x_k), t_k d_k \rangle + o(\|t_k d_k\|)$$

- In order to have $f(x_{k+1}) < f(x_k)$, we need

$$\langle \nabla f(x_k), d_k \rangle < 0$$

- The most natural choice is

$$d_k = -\nabla f(x_k)$$

- **Steepest descent**: $\boxed{x_{k+1} = x_k - t_k \nabla f(x_k)}$

Newton method

Quadratic approximation: Let $f : \mathbb{R}^N \to \mathbb{R}$ be **twice** continuously differentiable and $\mathbf{h}$ in $\mathbb{R}^N$.
Then, the best quadratic approximation of f in a neighbourhood of $\mathbf{x}$ is

$$T(\mathbf{x}, \mathbf{h}) = \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{h}, \nabla^2 f(\mathbf{x}) \mathbf{h} \rangle}_{m_k(\mathbf{x}, \mathbf{h})}$$

- When $\mathbf{h} = d_k$ and $\mathbf{x} = x_k$, we have:

$$f(x_{k+1}) \leq \underbrace{f(x_k) + \langle \nabla f(x_k), d_k \rangle + \langle d_k, \nabla^2 f(x_k) d_k \rangle}_{m_k(x_k, d_k)}$$

- The Newton direction is the minimizer of m_k is

$$d_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

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

Steepest descent

Quadratic approximation: Let $f : \mathbb{R}^N \to \mathbb{R}$ be $L-$smooth. Then for any $\beta \geq L$, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|^2$$

- When $\mathbf{y} = x_{k+1}$ and $\mathbf{x} = x_k$, we have:

$$f(x_{k+1}) \leq \underbrace{f(x_k) + \langle \nabla f(x_k), d_k \rangle + \frac{\beta}{2} \|d_k\|^2}_{m_k(x_k, d_k)}$$

- The step direction d_k leading to the minimum m_k is

$$d_k = -\frac{1}{\beta} \nabla f(x_k)$$

- Iterations: $\boxed{x_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k)}$

Newton method

Quadratic approximation: Let $f : \mathbb{R}^N \to \mathbb{R}$ be **twice** continuously differentiable. Then,

$$f(\mathbf{y}) = \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \langle \mathbf{y} - \mathbf{x}, H_k(\mathbf{y} - \mathbf{x}) \rangle}_{m_k(\mathbf{x}, \mathbf{y} - \mathbf{x})}$$

where H_k symmetric positive-definite.

- When $\mathbf{y} = x_{k+1}$ and $\mathbf{x} = x_k$, we have:

$$f(x_{k+1}) \leq \underbrace{f(x_k) + \langle \nabla f(x_k), d_k \rangle + \langle d_k, H_k d_k \rangle}_{m_k(x_k, d_k)}$$

- The Newton direction is the minimizer of m_k is

$$d_k = -H_k^{-1} \nabla f(x_k)$$

- Iterations: $\boxed{x_{k+1} = x_k - H_k^{-1} \nabla f(x_k)}$

Newton method

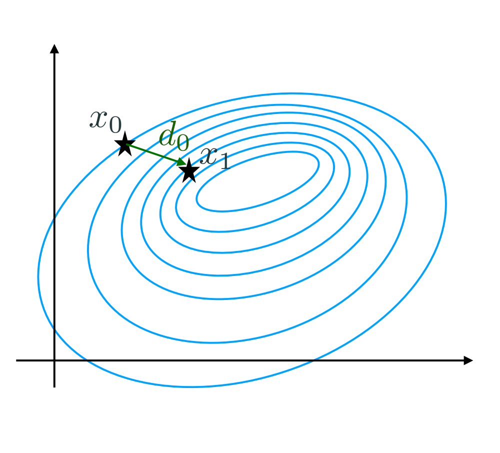
- Iterations: $\boxed{x_{k+1} = x_k - H_k^{-1} \nabla f(x_k)}$
- Choice for H_k:
  - $H_k = t_k \nabla^2 f(x_k)$
  - H_k diagonal e.g. $h_{k,k} = \frac{d^2 f(x_k)}{dx_k^2}$
  - $H_k = t_k \nabla^2 f(x_0)$

Step-size choice

- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.
- Iteration type :

  where

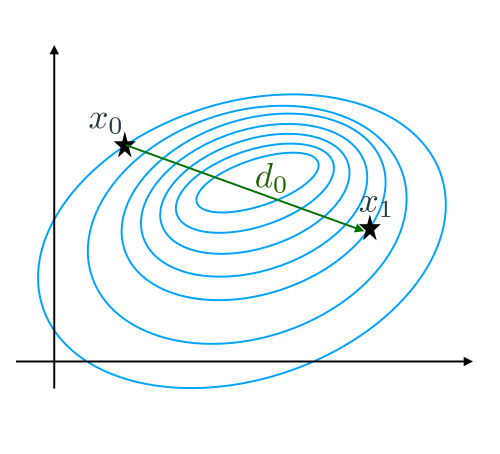
  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.



Step-size choice

- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.

- Iteration type :



where

- $t_k > 0$: step-length,
- $d_k \in \mathbb{R}^N$: step direction.

Step-size choice

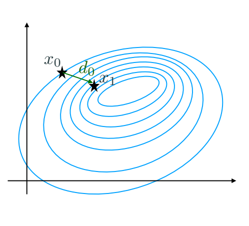
- Goal: build a sequence $(x_k)_{k\in\mathbb{N}}$ that converges to $\widehat{x}$.

- Iteration type :

$$x_{k+1} = x_k + t_k d_k$$

where

  - $t_k > 0$: step-length,
  - $d_k \in \mathbb{R}^N$: step direction.



- Choice of d_k (cf. previous slides)
- **Choice of t_k**
  - Armijo condition : not too large.
  - Wolfe condition : not too small.

Armijo condition

The **Armijo rule** requires that

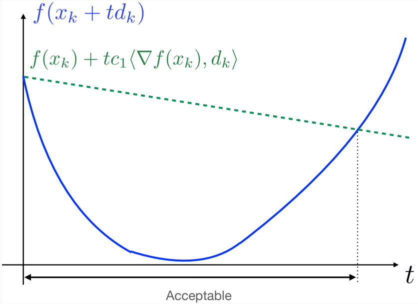
$$f(x_k + t_k d_k) \leq f(x_k) + t_k c_1 \langle \nabla f(x_k), d_k \rangle$$

where $c_1 \in (0, 1)$.

Remarks:

- Armijo rule is stronger than just asking the simple descrease
  $$f(x_{k+1}) \leq f(x_k)$$
  because
  $$\langle \nabla f(x_k), d_k \rangle < 0.$$

- Choosing t according to Armijo rule avoids choosing t too large.



Wolfe condition

The **Wolfe rule** requires that

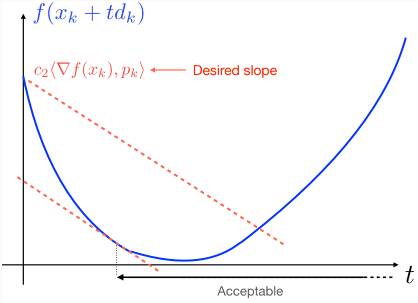
$$\langle \nabla f(x_k + t_k d_k), d_k \rangle \geq c_2 \langle \nabla f(x_k), d_k \rangle$$

where $c_2 \in (c_1, 1)$.

Remarks:

- Require that the slope of
  $$f(x_k + t_k d_k)$$
  to be greater than the negative slope $c_2 \langle \nabla f(x_k), d_k \rangle$

- Ensure that the slope has been reduced sufficiently.



Backtracking strategy

Algorithm: Given x_k, t_0, d_k, $b_{\max}$, $c_1 \in (0,1)$, $\gamma \in (0,1)$

$t_k = t_0$

For $b = 0, 1, \ldots, b_{\max}$

If $f(x_k + t_k d_k) \leq f(x_k) + t_k c_1 \langle \nabla f(x_k), d_k \rangle$ (satisfy Armijo)

Stop

Otherwise set $t_k = \gamma t_k$

Remarks:

- The name backtracking is due to the fact that k if progressively reduced.

BFGS method

- **BFGS**: Broyden, Fletcher, Goldfarb, and Shanno (1970)
- Quasi-Newton algorithm:

$$x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k)$$

  where B_k is a symmetric positive definite matrix that will be updated at each iteration.

BFGS method

- **BFGS**: Broyden, Fletcher, Goldfarb, and Shanno (1970)
- Quasi-Newton algorithm:

$$x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k)$$

  where B_k is a symmetric positive definite matrix that will be updated at each iteration.

- B_k is used in place of the true Hessian in the Newton method.
- **How to choose B_k ?**
  - Secant equation: $B_{k+1} s_k = y_k$
    where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

BFGS method

- **BFGS**: Broyden, Fletcher, Goldfarb, and Shanno (1970)
- Quasi-Newton algorithm:

$$x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k)$$

  where B_k is a symmetric positive definite matrix that will be updated at each iteration.
- B_k is used in place of the true Hessian in the Newton method.
- **How to choose B_k ?**
  - Secant equation: $B_{k+1} s_k = y_k$
    where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

    $\rightarrow$ interpretation: extension of the finite difference approximation of the second order derivative.
  - Secant equation for $B_{k+1}^{-1} = H_{k+1}$:

$$\boxed{H_{k+1} y_k = s_k}$$

BFGS method

- H_{k+1} should satisfy the secant equation and must be positive definite:

$$\min_H \|H - H_k\| \quad \text{s.t.} \quad \begin{cases} H = H^\top \\ H y_k = s_k \end{cases}$$

- The unique solution H_{k+1} when considering a weighted Frobenius norm is:

$$H_{k+1} = (\mathrm{Id} - \rho_k s_k y_k^\top) H_k (\mathrm{Id} - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top$$

  where

$$\begin{cases} s_k = x_{k+1} - x_k \\ y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \\ \rho_k = \frac{1}{\langle y_k, s_k \rangle} \end{cases}$$