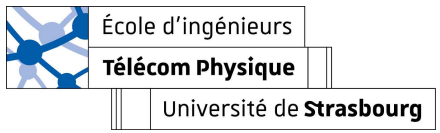
Optimization

Hassan OMRAN

Lecture 3: Multi-Dimensional Search Methods - Part I

Télécom Physique Strasbourg
Université de Strasbourg



Outline of the talk

1. Gradient methods

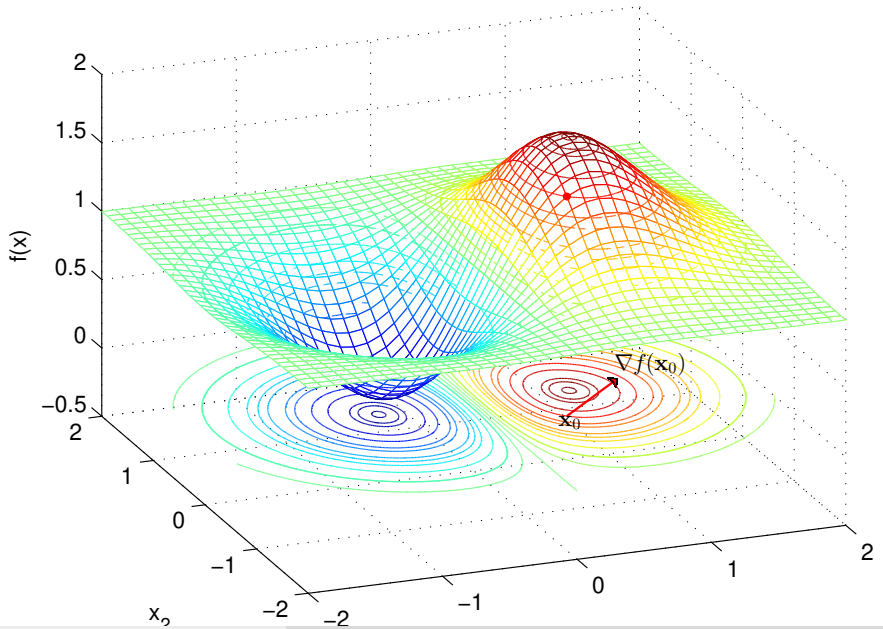
2. Newton's method

Gradient methods

Consider the continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. *We study the following problem:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

- *Remember that the gradient* $\nabla f(\boldsymbol{x}_0)$ *indicate the direction of the maximum increase of* $f(\cdot)$ *at* $\boldsymbol{x}_0$
- *The vector* $-\nabla f(\boldsymbol{x}_0)$ *indicates the direction of the* ***maximum decrease*** *of* $f(\cdot)$ *at* $\boldsymbol{x}_0$



Gradient methods

Starting from a point $\boldsymbol{x}_0$*, we look for a new point* $\boldsymbol{x}_1$ *in the direction* $-\nabla f(\boldsymbol{x}_0)$*. It is easy to show that*

$$f\Big(\boldsymbol{x}_0 - \alpha \nabla f(\boldsymbol{x}_0)\Big) = f(\boldsymbol{x}_0) - \alpha \|\nabla f(\boldsymbol{x}_0)\|^2 + o(\alpha)$$

Clearly when $\nabla f(\boldsymbol{x}_0) \neq \boldsymbol{0}$ *and for sufficiently small* $\alpha > 0$ *we have*

$$f\Big(\boldsymbol{x}_0 - \alpha \nabla f(\boldsymbol{x}_0)\Big) < f(\boldsymbol{x}_0)$$

Algorithm:

Stating from a point $\boldsymbol{x}_k$ *we calculate the gradient* $\nabla f(\boldsymbol{x}_k)$*, and for a suitable step size* α_k *we define the next point*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \tag{1}$$

The question is how to choose the step size α_k *?*

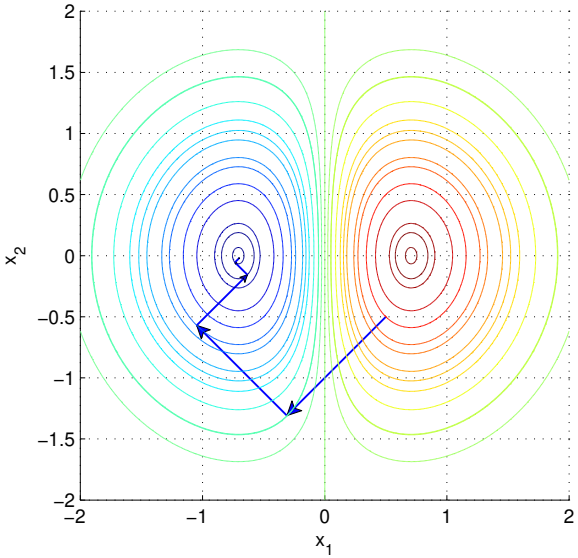
- *Gradient methods with a fixed step size* $\alpha_k = \alpha > 0$
- *Steepest descent*

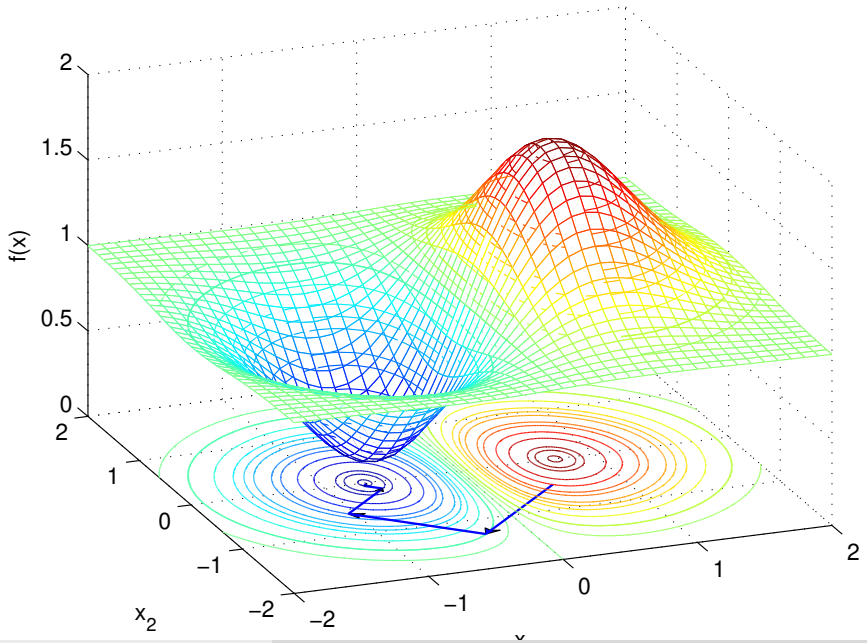
Steepest descent

The steepest descent is a gradient method where the step α_k is chosen in order to have the maximum amount of decrease of the objective function $f(\cdot)$ at each iteration. That is, at each iteration we define the function $h_k(\alpha) = f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big)$ and we solve

$$\alpha_k = \arg\min_{\alpha \geq 0} h_k(\alpha) = \arg\min_{\alpha \geq 0} f\big(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)\big) \tag{2}$$

Note that (2) is a one-dimensional optimization problem.





Sequence resulting from the steepest descent method.

Steepest descent: orthogonal generated directions

The next theorem shows that the directions generated by the steepest descent method are orthogonal.

Theorem 1

Let $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ *be the sequence from the steepest descent for a function* $f : \mathbb{R}^n \to \mathbb{R}$, *then* $\forall k \geq 0$ *the vector* $\boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}$ *is orthogonal to the vector* $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$.

Proof.

Consider the function
$h_k : \mathbb{R} \to \mathbb{R}$ *defined by* $h_k(\alpha) = f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))$. *From the FONC and the chain rule*

$$0 = \dot{h}_k(\alpha)\big|_{\alpha=\alpha_k} = \Big(\nabla f(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k))\Big)^T (-\nabla f(\boldsymbol{x}_k)) = -\langle \nabla f(\boldsymbol{x}_{k+1}), \nabla f(\boldsymbol{x}_k) \rangle \tag{3}$$

Finally, since $\boldsymbol{x}_{k+2} = \boldsymbol{x}_{k+1} - \alpha_{k+1} \nabla f(\boldsymbol{x}_{k+1})$ *and* $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$

$$\langle \boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle = \alpha_{k+1} \alpha_k \langle \nabla f(\boldsymbol{x}_{k+1}), \nabla f(\boldsymbol{x}_k) \rangle \tag{4}$$

From (3) *and* (4) *we have that* $\langle \boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle = 0$.

□

Steepest descent: descent property

At each iteration, if $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$, then the value of the function will strictly decrease

Theorem 2

Let $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ be the sequence from the steepest descent for a function $f : \mathbb{R}^n \to \mathbb{R}$. If $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$ then $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$.

Proof

Consider the functions $h_k : \mathbb{R} \to \mathbb{R}$ defined by $h_k(\alpha) = f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))$. Remember from (2) that

$$h_k(\alpha_k) \leq h_k(\alpha), \qquad \forall \alpha \geq 0 \tag{5}$$

Also

$$\dot{h}_k(0) = \Big(\nabla f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))\Big)^T \Big(- \nabla f(\boldsymbol{x}_k)\Big)\Big|_{\alpha=0} = -\|\nabla f(\boldsymbol{x}_k)\|^2 < 0$$

Thus, from Taylor theorem we have

$$h_k(\alpha) = h_k(0) + \dot{h}_k(0)\alpha + o(\alpha)$$

Steepest descent: descent property

Proof (Cont.)

then for sufficiently small $\overline{\alpha} > 0$ *we have*

$$h_k(\overline{\alpha}) < h_k(0) \tag{6}$$

From (5) *and* (6)

$$\begin{aligned} h_k(\alpha_k) \leq h_k(\overline{\alpha}) &< h_k(0) \\ f(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)) &< f(\boldsymbol{x}_k) \\ f(\boldsymbol{x}_{k+1}) &< f(\boldsymbol{x}_k) \end{aligned}$$

□

Steepest descent: stopping criteria

If at a certain iteration we have that $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$, then the point satisfies the FONC, and he algorithm can be stopped.

Practically, it is not always possible to obtain $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$. One of the following criteria can be used as a stopping condition (with a small $\epsilon > 0$):

- $\|\nabla f(\boldsymbol{x}_k)\| \leq \epsilon$
- $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| < \epsilon$
- $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \epsilon$
- $\frac{|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)|}{|f(\boldsymbol{x}_k)|} < \epsilon$
- $\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k\|} < \epsilon$

The last two criteria are scale-independent. They are to be used with caution as dividing by very small numbers may occur.

Steepest descent: the case of quadratic function

The case of quadratic function

Consider the following problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \\ s.t. \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned} \tag{7}$$

for a given matrix $0 \prec Q = Q^T \in \mathbb{R}^{n\times n}$ *and a vector* $\boldsymbol{q} \in \mathbb{R}^n$. *Note that* $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} + \boldsymbol{q}$ *and* $D^2 f(\boldsymbol{x}) = Q \succ 0$.

Steepest descent: the case of quadratic function

The case of quadratic function

Consider the following problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned} \tag{7}$$

for a given matrix $0 \prec Q = Q^T \in \mathbb{R}^{n\times n}$ *and a vector* $\boldsymbol{q} \in \mathbb{R}^n$. *Note that* $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} + \boldsymbol{q}$ *and* $D^2 f(\boldsymbol{x}) = Q \succ 0$.

From SOSC with $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$ *and* $D^2 f(\boldsymbol{x}) = Q \succ 0$ *we find that the solution is* $\boldsymbol{x}^\star = -Q^{-1}\boldsymbol{q}$.
In this case, we will show that we can provide the expression of α_k.

$$\begin{aligned} \alpha_k &= \arg\min_{\alpha\geq 0} h_k(\alpha) = \arg\min_{\alpha\geq 0} f\Big(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)\Big) \\ &= \arg\min_{\alpha\geq 0} \frac{1}{2}\Big(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)\Big)^T Q\Big(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)\Big) + \boldsymbol{q}^T\Big(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)\Big) \end{aligned}$$

From FONC $\dot{h}_k(\alpha)\big|_{\alpha=\alpha_k} = 0$ *we find that*

Steepest descent: the case of quadratic function

$$
\begin{aligned}
0 &= \dot{h}_k(\alpha)\Big|_{\alpha=\alpha_k} = \frac{d}{d\alpha} f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))\Big|_{\alpha=\alpha_k} \\
&= \Big(\nabla f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k))\Big)^T \Big(-\nabla f(\boldsymbol{x}_k)\Big)\Big|_{\alpha=\alpha_k} \\
&= \Big(Q(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)) + \boldsymbol{q}\Big)^T \Big(-\nabla f(\boldsymbol{x}_k)\Big)\Big|_{\alpha=\alpha_k} \\
&= \Big((Q\boldsymbol{x}_k + \boldsymbol{q}) - \alpha_k Q \nabla f(\boldsymbol{x}_k)\Big)^T \Big(-\nabla f(\boldsymbol{x}_k)\Big)
\end{aligned}
$$

we suppose that $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$*, since if it is the case, then* $\boldsymbol{x}_k = \boldsymbol{x}^\star$ *and the algorithm stops*

$$
\Rightarrow \alpha_k = \frac{(Q\boldsymbol{x}_k + \boldsymbol{q})^T \nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_k)^T Q \nabla f(\boldsymbol{x}_k)} = \frac{\nabla f(\boldsymbol{x}_k)^T \nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_k)^T Q \nabla f(\boldsymbol{x}_k)} \tag{8}
$$

finally

$$
\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\nabla f(\boldsymbol{x}_k)^T \nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_k)^T Q \nabla f(\boldsymbol{x}_k)} \nabla f(\boldsymbol{x}_k) \tag{9}
$$

Steepest descent: the case of quadratic function

In the case of quadratic function, it is shown that the steepest descent always converges.

Theorem 3

The descent method in the case of quadratic function in (7) *converges for any initial condition. That is* $\boldsymbol{x}_k \rightarrow \boldsymbol{x}^\star \ \forall \boldsymbol{x}_0 \in \mathbb{R}^n$.

Also, in this case the convergence rate is linear.

Theorem 4

The descent method in the case of quadratic function in (7) *has a linear convergence*

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}^\star\| \sqrt{\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} - 1}$$

the ratio $\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} \geq 1$ *is called the* ***condition number*** *of* Q. *The smaller this number, the faster* $\boldsymbol{x}_k$ *converges to* $\boldsymbol{x}^\star$.

Steepest descent: the case of quadratic function

Proof

Consider the error $e_k = \boldsymbol{x}_k - \boldsymbol{x}^\star$, *and the function* $V(e_k) = e_k^T Q e_k$. *For simplicity, we will use the notation* $\boldsymbol{g}_k$ *for* $\nabla f(\boldsymbol{x}_k)$.

$$\begin{aligned} V(e_{k+1}) &= (\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star)^T Q(\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star) \\ &= (\boldsymbol{x}_k - \boldsymbol{x}^\star - \alpha_k \boldsymbol{g}_k)^T Q(\boldsymbol{x}_k - \boldsymbol{x}^\star - \alpha_k \boldsymbol{g}_k) \\ &= (\boldsymbol{x}_k - \boldsymbol{x}^\star)^T Q(\boldsymbol{x}_k - \boldsymbol{x}^\star) - \alpha_k \boldsymbol{g}_k^T Q(\boldsymbol{x}_k - \boldsymbol{x}^\star) - \alpha_k (\boldsymbol{x}_k - \boldsymbol{x}^\star)^T Q \boldsymbol{g}_k + \alpha_k^2 \boldsymbol{g}_k^T Q \boldsymbol{g}_k \\ &= V(e_k) - 2\alpha_k \boldsymbol{g}_k^T \underbrace{Q(\boldsymbol{x}_k - \boldsymbol{x}^\star)}_{Q\boldsymbol{x}_k + \boldsymbol{q} = \boldsymbol{g}_k} + \alpha_k^2 \boldsymbol{g}_k^T Q \boldsymbol{g}_k \end{aligned}$$

From (8) *we have*

$$\begin{aligned} V(e_{k+1}) &= V(e_k) - 2\Big(\frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T Q \boldsymbol{g}_k}\Big)\boldsymbol{g}_k^T \boldsymbol{g}_k + \Big(\frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T Q \boldsymbol{g}_k}\Big)^2 \boldsymbol{g}_k^T Q \boldsymbol{g}_k \\ &= V(e_k) - \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T Q \boldsymbol{g}_k} \end{aligned}$$

Steepest descent: the case of quadratic function

Proof (Cont.)

$$\frac{V(e_{k+1})}{V(e_k)} = 1 - \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T Q \boldsymbol{g}_k} \frac{1}{V(e_k)}$$

Note that

$$\begin{aligned} V(e_k) &= e_k^T Q e_k \\ &= (\boldsymbol{x}_k - \boldsymbol{x}^\star)^T Q (\boldsymbol{x}_k - \boldsymbol{x}^\star) \\ &= (Q\boldsymbol{x}_k - Q\boldsymbol{x}^\star)^T Q^{-1} (Q\boldsymbol{x}_k - Q\boldsymbol{x}^\star) \\ &= \boldsymbol{g}_k^T Q^{-1} \boldsymbol{g}_k \end{aligned}$$

Thus

$$\frac{V(e_{k+1})}{V(e_k)} = 1 - \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{(\boldsymbol{g}_k^T Q \boldsymbol{g}_k)(\boldsymbol{g}_k^T Q^{-1} \boldsymbol{g}_k)}$$

Steepest descent: the case of quadratic function

Proof (Cont.)

From Rayleigh's inequality we have

$$\frac{V(e_{k+1})}{V(e_k)} = 1 - \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{(\boldsymbol{g}_k^T Q \boldsymbol{g}_k)(\boldsymbol{g}_k^T Q^{-1} \boldsymbol{g}_k)} \leq 1 - \frac{1}{\lambda_{max}(Q)\lambda_{max}(Q^{-1})}$$

$$\frac{\lambda_{min}(Q)\|e_{k+1}\|^2}{\lambda_{max}(Q)\|e_k\|^2} \leq \frac{V(e_{k+1})}{V(e_k)} \leq 1 - \frac{\lambda_{min}(Q)}{\lambda_{max}(Q)}$$

finally

$$\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|^2}{\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2} \leq \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} - 1$$

□

Steepest descent: the case of quadratic function

Examples

Apply the steepest descent method to the problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

in the following cases

- $$Q = \lambda I,\ \lambda > 0, \qquad \forall \boldsymbol{q} \in \mathbb{R}^2 \qquad \forall \boldsymbol{x}_0 \in \mathbb{R}^2$$
- $$Q = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{q} = \begin{bmatrix} -3 \\ -3 \end{bmatrix}, \qquad \boldsymbol{x}_0 = \begin{bmatrix} -2 \\ -7 \end{bmatrix}$$
- $$Q = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{q} = \begin{bmatrix} -3 \\ -3 \end{bmatrix}, \qquad \boldsymbol{x}_0 = \begin{bmatrix} -2 \\ -7 \end{bmatrix}$$

Steepest descent: remarks

- *Gradient methods are simple, and have (often) good performance*
- *They require the calculation of the gradient, which may not be always possible*
- *They have good convergence properties for convex QP problems*
- *They may have bad performance for some functions*

  We will see how, for the case of Rosenbock's function, they have a slow zig-zag convergence near the optimal point, and require a big number of function evaluations.

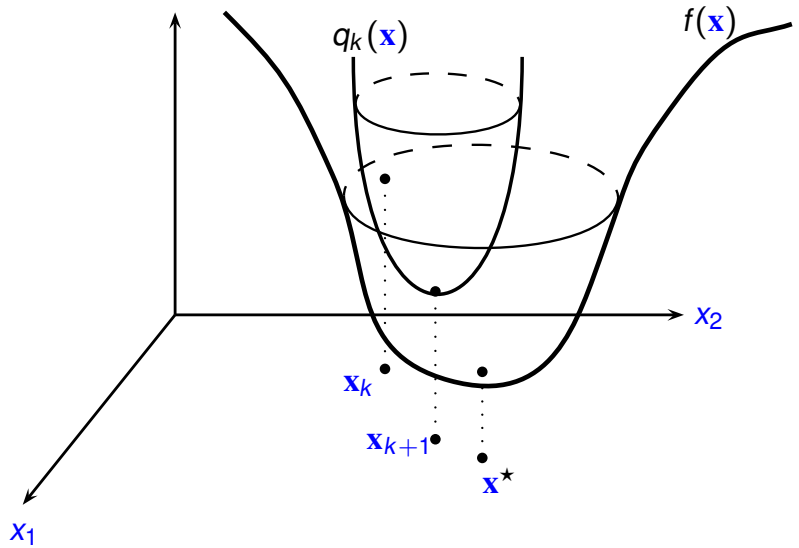
Newton's method

*Consider the **twice** continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we study the following problem:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned} \tag{10}$$

The idea is to minimize a second-order approximation of $f(\cdot)$ at each iteration

- *Start with an initial value $\boldsymbol{x}_0$*
- *At each iteration minimize second-order approximation of $f(\cdot)$ at $\boldsymbol{x}_k$.*
- *Use the minimizer of the quadratic approximate as a starting point for the next iteration*



The approximation of the function $f(\cdot)$ by a quadratic function $q_k(\cdot)$.

Newton's method

At iteration k, using Taylor series expansion of $f(\cdot)$ at $\boldsymbol{x}_k$, we find the approximate $q_k(\cdot)$

$$q_k(\boldsymbol{x}) = f(\boldsymbol{x}_k) + (\boldsymbol{x} - \boldsymbol{x}_k)^T \nabla f(\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^T D^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) \tag{11}$$

Note that

$$\nabla q_k(\boldsymbol{x}) = \nabla f(\boldsymbol{x}_k) + D^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k), \qquad D^2 q_k(\boldsymbol{x}) = D^2 f(\boldsymbol{x}_k)$$

If $D^2 f(\boldsymbol{x}_k) > 0$ then $q_k(\cdot)$ achieves a minimum for $\nabla q_k(\boldsymbol{x}_{k+1}) = \boldsymbol{0}$, that is

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \left(D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k) \tag{12}$$

Denoting $A = D^2 f(\boldsymbol{x}_k)$ and $b = \nabla f(\boldsymbol{x}_k)$, the recursive formula is composed of two steps:

- *solving $Ay = b$*
- $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - y$

Newton's method: convergence rate

The case of quadratic function

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \\ s.t. & \boldsymbol{x} \in \mathbb{R}^n \end{array} \tag{13}$$

For a given matrix $0 \prec Q = Q^T \in \mathbb{R}^{n\times n}$ *and a vector* $q \in \mathbb{R}^n$. *Note that* $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} + \boldsymbol{q}$ *and* $D^2 f(\boldsymbol{x}^\star) = Q \succ 0$.

Starting from any initial point $\boldsymbol{x}_0$, *we have*

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{x}_0 - Q^{-1}(Q\boldsymbol{x}_0 + \boldsymbol{q}) \\ &= -Q^{-1}\boldsymbol{q} \\ &= \boldsymbol{x}^\star \end{aligned}$$

In this case, the Newton's method find the solution in one iteration, starting from any initial condition.

Newton's method: convergence rate

In the general case, the method has good convergence properties when starting close the solution.

Theorem 5

Consider the problem in (10) *with three times continuously differentiable* $f(\cdot)$*, and a point* $\boldsymbol{x}^\star$ *with* $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$ *and invertible* $D^2 f(\boldsymbol{x}^\star)$*. Then, Newton's method is well defined* $\forall k \geq 0$*, and converges to* $\boldsymbol{x}^\star$ *if the starting point* $\boldsymbol{x}_0$ *is sufficiently close to* $\boldsymbol{x}^\star$*.*
Moreover, it has a quadratic convergence

$$\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2} \leq c, \qquad c > 0.$$

Note that theorem above shows the ***local*** *convergence property of the method.*

If the Hessian $D^2 f(\boldsymbol{x}_k)$ *is not positive definite, the method may not have the descent property.*

Newton's method: convergence rate

The following theorem shows that the algorithm generates decreasing directions when $D^2 f(\boldsymbol{x}_k) \succ 0$.

Theorem 6

If the Hessian $D^2 f(\boldsymbol{x}_k) \succ 0$ *and* $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$, *then the direction generated by Newton's method*

$$\boldsymbol{d} = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\left(D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k)$$

is a descent direction.

Proof

Consider the function $h_k(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d})$. *Using Taylor's theorem we have*

$$h_k(\alpha) = h_k(0) + \underbrace{\dot{h}_k(0)}_{\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d} = \boldsymbol{d}^T \nabla f(\boldsymbol{x}_k) = -\boldsymbol{d}^T \left(D^2 f(\boldsymbol{x}_k)\right) \boldsymbol{d}} \alpha + o(\alpha)$$

Newton's method: convergence rate

Proof (Cont.)

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}) \quad = \quad f(\boldsymbol{x}_k) - \alpha \boldsymbol{d}^T \Big(D^2 f(\boldsymbol{x}_k)\Big) \boldsymbol{d} + o(\alpha)$$

Since $\boldsymbol{d} \neq \boldsymbol{0}$, *and* $D^2 f(\boldsymbol{x}_k) \succ 0$, *then* $\exists\, \overline{\alpha} > 0$ *such that*

$$f\Big(\boldsymbol{x}_k + \alpha \boldsymbol{d}\Big) < f(\boldsymbol{x}_k), \qquad \forall \alpha \in (0, \overline{\alpha})$$

□

Newton's method: modifications

By Theorem 6 we see that instead of $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \left(D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k)$, *we can introduce the following one-dimensional optimization*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \left(D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k)$$

with

$$\alpha_k = \arg\min_{\alpha \geq 0} f\left(\boldsymbol{x}_k - \alpha \left(D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k)\right) \tag{14}$$

Note that by Theorem 6 the method in (14) *guarantees the descent property* $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ *(when* $D^2 f(\boldsymbol{x}_k) \succ 0$ *and* $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$*).*

Finally, the ***Levenberg–Marquardt*** *method is a modification of Newton's method which overcomes the problem caused by non definite positive Hessians*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \left(\mu_k I + D^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k), \qquad \text{with } \mu_k \geq 0$$

This can be seen as a combination of Newton's method when $\mu_k \to 0$, *and gradient methods* $\mu_k \to \infty$.

Newton's method: remarks

- *The method is also called Newton-Raphson method*
- *Can also be seen as a method to find* $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$
- *The quadratic approximate matches the first and second derivatives of* $f(\cdot)$ *at* $\boldsymbol{x}_k$*, thus more information are used than the gradient methods*
- *The method performs better than the steepest descent which uses only the first derivative (if the initial point is close the the minimizer)*
- *Drawbacks:*
  - *The evaluation of the Hessian can be computationally difficult*
  - *Requires solving a set of linear equations, which is of* $O(n^3)$ *complexity*
  - *Only local convergence is guaranteed*
  - *Modifications are needed if the Hessian is not definite positive*