Optimization

Màster de Fonaments de Ciència de Dades

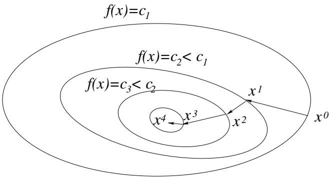
Lecture IV. Alternating directions methods for unconstrained optimization

Gerard Gómez

Alternating directions methods

- The main purpose of the alternating directions methods is to accelerate the convergence of the descent methods and, in this way, reduce the total number of iterations.

- In the alternating directions methods, we start at a certain starting position $\boldsymbol{x}$, along a direction $\boldsymbol{d}$, and then minimize $f(\boldsymbol{x} + \alpha \boldsymbol{d})$ selecting the suitable value of α



- Next we use $\boldsymbol{x} + \alpha^* \boldsymbol{d}$ as the new starting position, choose a different direction, and minimize along that direction.......

- Consequently, the basic tool for alternating directions methods, such as the gradient methods, is a 1-D minimization (Golden section, Fibonacci,...)

- Different alternating directions methods differ as to how the directions are chosen

Alternating directions methods. Contents

- First example: The coordinate descent method
- New definition: Conjugate directions
- Alternating directions methods:
  1. Conjugate gradient methods
  2. Conjugate gradient methods for quadratic functions
  3. Conjugate gradient methods for $\mathcal{C}^1$ functions
  4. Powell's method for continuous functions
  5. Powell's method for quadratic functions

Alternating directions methods. First example

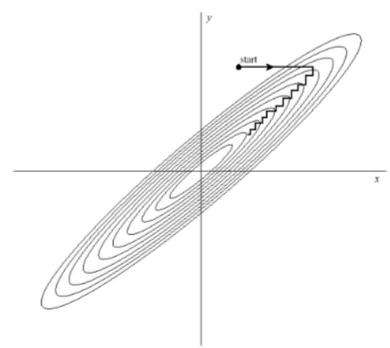
The coordinate descent method

- Use n orthogonal unit vectors in turn:

$$\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n, \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n, \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots$$

  as directions $\boldsymbol{d}$, and for each $\boldsymbol{e}_i$ minimize $f(\boldsymbol{x} + \alpha \boldsymbol{e}_i)$

- This method has slow convergence, unless the unit vectors are well-oriented with respect to the "valley" in which there is $f(\boldsymbol{x}^*)$



The steepest descent method

- We have already seen, in the steepest descent method the direction is given by the unitary vector

$$\boldsymbol{d}^{k+1} = -\frac{\nabla f(\boldsymbol{x}^k)}{\|\nabla f(\boldsymbol{x}^k)\|}$$

- Note that with this procedure we always choose a new direction that is orthogonal to the previous direction:

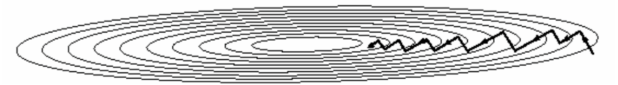
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha \nabla f(\boldsymbol{x}^k), \quad \Rightarrow$$

$$0 = \frac{df(x^{k+1})}{d\alpha} = \nabla f(\boldsymbol{x}^{k+1})^T \frac{d\boldsymbol{x}^{k+1}}{d\alpha} = -\nabla f(\boldsymbol{x}^{k+1})^T \nabla f(\boldsymbol{x}^k)$$

so

$$(\boldsymbol{d}^{k+1})^T \boldsymbol{d}^k = 0$$

- The performance isn't that good, because we can only ever take a right angle turn



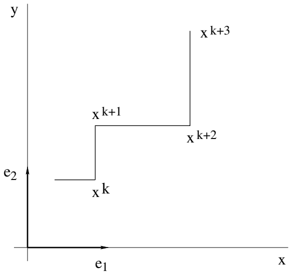
Alternating directions methods. Motivation

- Suppose that we are dealing with a 2-D problem, and that step k occurred along the y-axis, and led to position $\boldsymbol{x}^{k+1}$, at which

$$\frac{\partial f(\boldsymbol{x}^{k+1})}{\partial y} = 0$$

- The next step is along the x-axis: that step leads to a position $\boldsymbol{x}^{k+2}$, at which

$$\frac{\partial f(\boldsymbol{x}^{k+2})}{\partial x} = 0$$



Alternating directions methods. Motivation

- But (exercise) if

$$\frac{\partial^2 f(\mathbf{x}^{k+2})}{\partial y \partial x} \neq 0 \quad \Rightarrow \quad \frac{\partial f(\mathbf{x}^{k+2})}{\partial y} \neq 0$$

- We really want to move along some direction other than the x-axis, such that

$$\frac{\partial f(\mathbf{x}^{k+2})}{\partial y} = 0$$

- Thus the optimum direction is not along $\nabla f = \left(\frac{\partial f(\mathbf{x})}{\partial x}, \frac{\partial f(\mathbf{x})}{\partial y}\right)$ but rather in a direction that preserves the minimization achieved in the previous step (and, in multi-dimensions, all previous steps)

- Let us see how we can define these (conjugate) directions

Alternating directions methods. Conjugate directions

- Let $\boldsymbol{x}^k$, $\boldsymbol{x}^{k+1}$ and $\boldsymbol{x}^{k+2}$ be three consecutive points such that
  - $\boldsymbol{x}^{k+1}$ is the minimum of f along $\boldsymbol{x}^k + \lambda \boldsymbol{d}^k$, where

$$\boldsymbol{d}^k = \frac{\boldsymbol{x}^{k+1} - \boldsymbol{x}^k}{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|}$$

    and so

$$D_{\boldsymbol{d}^k} f(\boldsymbol{x}^{k+1}) = \nabla f(\boldsymbol{x}^{k+1})^T \boldsymbol{d}^k = 0$$

  - $\boldsymbol{x}^{k+2}$ is the minimum along $\boldsymbol{x}^{k+1} + \lambda \boldsymbol{d}^{k+1}$, where

$$\boldsymbol{d}^{k+1} = \frac{\boldsymbol{x}^{k+2} - \boldsymbol{x}^{k+1}}{\|\boldsymbol{x}^{k+2} - \boldsymbol{x}^{k+1}\|}$$

    and so

$$D_{\boldsymbol{d}^{k+1}} f(\boldsymbol{x}^{k+2}) = \nabla f(\boldsymbol{x}^{k+2})^T \boldsymbol{d}^{k+1} = 0$$

- In addition, we would also like that

$$D_{\boldsymbol{d}^k} f(\boldsymbol{x}^{k+2}) = \nabla f(\boldsymbol{x}^{k+2})^T \boldsymbol{d}^k = 0$$

Alternating directions methods. Conjugate directions

- To set this condition, consider the Taylor expansion

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{\delta} + \dots$$

- Taking the gradient of the Taylor expansion, we obtain

$$\nabla f(\boldsymbol{x}+\boldsymbol{\delta}) = \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})\boldsymbol{\delta} + \dots$$

- Since we want that

$$D_{\boldsymbol{d}^k} f(\boldsymbol{x}^{k+2}) = \nabla f(\boldsymbol{x}^{k+2})^T \boldsymbol{d}^k = (\boldsymbol{d}^k)^T \nabla f(\boldsymbol{x}^{k+2}) = 0,$$

  and using the above Taylor expansion

$$\nabla f(\boldsymbol{x}^{k+2}) = \nabla f(\boldsymbol{x}^{k+1} + \boldsymbol{d}^{k+1}) = \nabla f(\boldsymbol{x}^{k+1}) + \nabla^2 f(\boldsymbol{x}^{k+1})\boldsymbol{d}^{k+1} + \dots$$

  the condition $D_{\boldsymbol{d}^k} f(\boldsymbol{x}^{k+2}) = 0$ requires

$$(\boldsymbol{d}^k)^T \left[\nabla f(\boldsymbol{x}^{k+1}) + \nabla^2 f(\boldsymbol{x}^{k+1})\boldsymbol{d}^{k+1}\right] \approx 0$$

Alternating directions methods. Conjugate directions

- Since $(\boldsymbol{d}^k)^T \nabla f(\boldsymbol{x}^{k+1}) = \nabla f(\boldsymbol{x}^{k+1})^T \boldsymbol{d}^k = 0$, because $\boldsymbol{x}^{k+1}$ was obtained by minimizing f along the $\boldsymbol{d}^k$ (see page 8), it follows that the above condition

$$(\boldsymbol{d}^k)^T \left[\nabla f(\boldsymbol{x}^{k+1}) + \nabla^2 f(\boldsymbol{x}^{k+1}) \boldsymbol{d}^{k+1} \right] \approx 0$$

  becomes

$$\boxed{(\boldsymbol{d}^k)^T \nabla^2 f(\boldsymbol{x}^{k+1}) \boldsymbol{d}^{k+1} = 0}$$

- Definition

  If this last condition holds, we will say that

  $\boldsymbol{d}^k$ and $\boldsymbol{d}^{k+1}$ are conjugate with respect to $\nabla^2 f(\boldsymbol{x}^{k+1})$

- Clearly, this is different from steepest descent method, for which $(\boldsymbol{d}^k)^T \boldsymbol{d}^{k+1} = 0$

Alternating directions methods. Conjugate directions w.r.t. A

- One basic idea for alternating directions methods is the one related to conjugate directions which is a generalization of orthogonality

- Definition

  Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are said to be conjugate directions with respect to the $n \times n$ symmetric positive definite matrix A if

$$\boxed{\mathbf{x}^T A \mathbf{y} = 0}$$

- If A is symmetric positive definite matrix, then
  - It is well known that A has n orthogonal eigenvectors
  - These n vectors are also mutually conjugate, since

$$\mathbf{x}^T A \mathbf{y} = \mathbf{x}^T \lambda \mathbf{y} = \lambda \mathbf{x}^T \mathbf{y} = 0$$

Thus, for every $n \times n$ symmetric positive definite matrix there is at least one set of n mutually conjugate directions w.r.t. A

Conjugate directions

- **Remark**

  Let $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m$ $(m \leq n)$ be m nonzero vectors mutually conjugate with respect to A, then these vectors are linearly independent

  If this was not the case, then we could write

  $$\boldsymbol{d}_m = \sum_{i=1}^{m-1} \alpha_i \boldsymbol{d}_i$$

  from which it follows that

  $$(\boldsymbol{d}_m)^T A\, \boldsymbol{d}_m = 0$$

  that contradics the fact that $\boldsymbol{d}_m \neq 0$ and that A is positive definite

Conjugate directions. Construction

- Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ be k linearly independent vectors, then we can construct k mutually conjugate directions $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k$, with respect to A, such that

$$< \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k > = < \boldsymbol{d}_1, \ldots, \boldsymbol{d}_k >$$

The construction is similar to the Gram-Schmidt orthogonalization method. Since $\boldsymbol{d}_m^T A \boldsymbol{d}_m \neq 0$ (A is positive definite), we can define

$$\begin{aligned}
\boldsymbol{d}_1 &= \boldsymbol{v}_1 \\
\boldsymbol{d}_2 &= \boldsymbol{v}_2 - \frac{\boldsymbol{v}_2^T A \boldsymbol{d}_1}{\boldsymbol{d}_1^T A \boldsymbol{d}_1} \boldsymbol{d}_1 \\
\boldsymbol{d}_3 &= \boldsymbol{v}_3 - \frac{\boldsymbol{v}_3^T A \boldsymbol{d}_1}{\boldsymbol{d}_1^T A \boldsymbol{d}_1} \boldsymbol{d}_1 - \frac{\boldsymbol{v}_3^T A \boldsymbol{d}_2}{\boldsymbol{d}_2^T A \boldsymbol{d}_2} \boldsymbol{d}_2 \\
\vdots & \quad \vdots \\
\boldsymbol{d}_{i+1} &= \boldsymbol{v}_{i+1} - \sum_{m=1}^{i} \frac{\boldsymbol{v}_{i+1}^T A \boldsymbol{d}_m}{\boldsymbol{d}_m^T A \boldsymbol{d}_m} \boldsymbol{d}_m, \quad i = 3, \ldots, k-1
\end{aligned}$$

Clearly

$$\boldsymbol{v}_{i+1} \in < \boldsymbol{d}_1, \ldots, \boldsymbol{d}_{i+1} > \quad \text{and} \quad \boldsymbol{d}_{i+1} \in < \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i+1} >$$

so $< \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i+1} > = < \boldsymbol{d}_1, \ldots, \boldsymbol{d}_{i+1} >$ for $i = 1, \ldots, k-1$

Conjugate directions. Construction

- Now we need to proof that if $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_i$ are mutually conjugate w.r.t. A, then $\boldsymbol{d}_{i+1}^T A \boldsymbol{d}_j = 0$ for $j = 1, \ldots, i$

$$\boldsymbol{d}_{i+1}^T A \boldsymbol{d}_j = \boldsymbol{v}_{i+1}^T A \boldsymbol{d}_j - \sum_{m=1}^{i} \frac{\boldsymbol{v}_{i+1}^T A \boldsymbol{d}_m}{\boldsymbol{d}_m^T A \boldsymbol{d}_m} \boldsymbol{d}_m^T A \boldsymbol{d}_j = \boldsymbol{v}_{i+1}^T A \boldsymbol{d}_j - \frac{\boldsymbol{v}_{i+1}^T A \boldsymbol{d}_j}{\boldsymbol{d}_j^T A \boldsymbol{d}_j} \boldsymbol{d}_j^T A \boldsymbol{d}_j = 0$$

since $\boldsymbol{d}_m^T A \boldsymbol{d}_j = 0$ except if $m = 1$

Conjugate directions. Geometric interpretation

A geometric interpretation of conjugate vectors is the following. Let

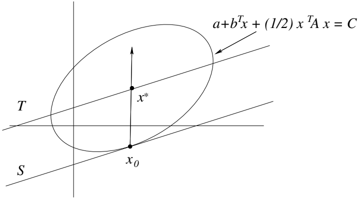
$$f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}$$

with A a symmetric positive definite matrix, be a quadratic function with a global minimum at $\boldsymbol{x}^*$

$$\nabla f(\boldsymbol{x}^*) = 0 \quad \Rightarrow \quad \boldsymbol{b} + A\boldsymbol{x}^* = 0 \quad \Rightarrow \quad \boldsymbol{x}^* = -A^{-1}\boldsymbol{b}$$

Then, the surfaces $f(\boldsymbol{x}) =$ constant are, generally, ellipsoids with center at $\boldsymbol{x}^*$

Let $\boldsymbol{x}_0$ be a point satisfying $f(\boldsymbol{x}_0) = c$



Then, we are going to see that the vector joining $\boldsymbol{x}_0$ and $\boldsymbol{x}^*$ is conjugate with respect to A to every vector in the tangent hyperplane to the ellipsoid at $\boldsymbol{x}_0$

Conjugate directions. Geometric interpretation

Definition

Given a point $\mathbf{x}_0 \in \mathbb{R}^n$*, the set of points satisfying*

$$\mathbf{x} = \mathbf{x}_0 + \sum_{j=1}^{m} \alpha_j \mathbf{z}^j$$

where the $\mathbf{z}^j$ *are* m *linearly independent vectors, and the* α_j *are arbitrary numbers, is an affine space or linear manifold generated by* $\mathbf{x}_0$ *and* $\mathbf{z}^1, \dots, \mathbf{z}^m$

Definition

Two affine spaces S *and* T $(S \neq T)$ *are parallel if they are generated by the same set of vectors* $\mathbf{z}_1, \dots, \mathbf{z}_m$ *but at different points:* $\mathbf{x}(S) \in S$, $\mathbf{x}(T) \in T$*, and* $\mathbf{x}(S) \neq \mathbf{x}(T)$

$$S = \left\{ \mathbf{x} \,\middle|\, \mathbf{x} = \mathbf{x}(S) + \sum_{j=1}^{m} \alpha_j \mathbf{z}^j \right\}, \quad T = \left\{ \mathbf{x} \,\middle|\, \mathbf{x} = \mathbf{x}(T) + \sum_{j=1}^{m} \alpha_j \mathbf{z}^j \right\}$$

Conjugate directions. Geometric interpretation

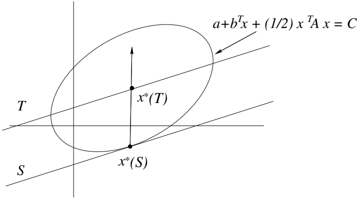
Theorem

Let $\mathbf{x}^*(S)$ and $\mathbf{x}^*(T)$ be the points that minimize

$$f(\mathbf{x}) = \mathbf{a} + \mathbf{b}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T A\mathbf{x},$$

with A a symmetric positive definite matrix, in two parallel affine spaces S and T. Then $\mathbf{x}^*(S) - \mathbf{x}^*(T)$ and any direction $\mathbf{z}$ contained in S and T are conjugate w.r.t. A, this is

$$\mathbf{z}^T A[\mathbf{x}^*(S) - \mathbf{x}^*(T)] = 0$$



Conjugate directions. Geometric interpretation

Proof:

According to the definition of $f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}$

$$\frac{d}{d\alpha}[f(\boldsymbol{x}^*(S) + \alpha\boldsymbol{z})] =$$
$$= \frac{d}{d\alpha}\left[\boldsymbol{a} + \boldsymbol{b}^T x^*(S) + \alpha\boldsymbol{b}^T z + \frac{1}{2}\left((\boldsymbol{x}^*(S) + \alpha\boldsymbol{z})^T A(\boldsymbol{x}^*(S) + \alpha\boldsymbol{z})\right)\right] =$$
$$= \boldsymbol{b}^T z + (x^*(S))^T A\boldsymbol{z} + \alpha\boldsymbol{z}^T A\boldsymbol{z}$$

Let $\boldsymbol{z}$ be a direction of S and T. According to the above computation:

$$\frac{d}{d\alpha}[f(\boldsymbol{x}^*(S) + \alpha\boldsymbol{z})]_{\alpha=0} = 0 \quad \Rightarrow \quad \boldsymbol{z}^T[A\boldsymbol{x}^*(S) + \boldsymbol{b}] = 0$$
$$\frac{d}{d\alpha}[f(\boldsymbol{x}^*(T) + \alpha\boldsymbol{z})]_{\alpha=0} = 0 \quad \Rightarrow \quad \boldsymbol{z}^T[A\boldsymbol{x}^*(T) + \boldsymbol{b}] = 0$$

so

$$\boldsymbol{z}^T A[\boldsymbol{x}^*(S) - \boldsymbol{x}^*(T)] = 0$$

□

Conjugate directions

Theorem

Let $\boldsymbol{z}_1, ..., \boldsymbol{z}_m$ such that $\boldsymbol{z}_i \in \mathbb{R}^n$, $\boldsymbol{z}_i \neq 0$, $m \leq n$, and that they are m mutually conjugate directions with respect to the symmetric poisitive definite matrix A, then the minimum of the quadratic function

$$f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T \boldsymbol{x} + \frac{1}{2} \boldsymbol{x}^T A \boldsymbol{x}$$

over the affine set generated by the point $\boldsymbol{x}_0 \in \mathbb{R}^n$ and the vectors $\boldsymbol{z}_1, ..., \boldsymbol{z}_m$ will be found by ***searching along each of the conjugate directions only once***

Conjugate directions. Proof of the theorem

Proof: The minimum will be a point $\boldsymbol{x}_0 + \alpha_1^* \boldsymbol{z}_1 + \ldots + \alpha_m^* \boldsymbol{z}_m$, such that the α_i^* minimize

$$f\left(\boldsymbol{x}_0 + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j\right) = \boldsymbol{a} + \boldsymbol{b}^T \left(\boldsymbol{x}_0 + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j\right) + \frac{1}{2}\left(\boldsymbol{x}_0 + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j\right)^T A \left(\boldsymbol{x}_0 + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j\right) =$$

$$= f(\boldsymbol{x}_0) + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j^T \boldsymbol{b} + \sum_{j=1}^{m} \alpha_j \boldsymbol{z}_j^T A \boldsymbol{x}_0 + \frac{1}{2} \sum_{j=1}^{m} \alpha_j^2 \boldsymbol{z}_j^T A \boldsymbol{z}_j =$$

$$= f(\boldsymbol{x}_0) + \sum_{j=1}^{m} \left[\alpha_j \boldsymbol{z}_j^T (\boldsymbol{b} + A\boldsymbol{x}_0) + \frac{1}{2}\alpha_j^2 z_j^T A \boldsymbol{z}_j\right]$$

Since in the last expression there are no $\alpha_j \alpha_k$ terms with $j \neq k$, the optimal α_j are found minimizing each summand:

$$\min_{\alpha_j} \left[f(\boldsymbol{x}_0) + \alpha_j \boldsymbol{z}_j^T (\boldsymbol{b} + A\boldsymbol{x}_0) + \frac{1}{2}\alpha_j^2 \boldsymbol{z}_j^T A \boldsymbol{z}_j\right] = \min_{\alpha_j} f(\boldsymbol{x}_0 + \alpha_j \boldsymbol{z}_j), \quad j = 1, \ldots, m$$

$\square$

Conjugate directions. Example

Example. Consider the quadratic function

$$f(x,y) = 2x^2 + 6y^2 + 2xy + 2x + 3y + 3$$

that can also be written as

$$f(x,y) = \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 12 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + 3$$

We choose $\boldsymbol{z}_1 = (1,0)^T$. A conjugate direction to $\boldsymbol{z}_1$ with respect to

$$A = \begin{pmatrix} 4 & 2 \\ 2 & 12 \end{pmatrix}$$

is $\boldsymbol{z}_2 = (-1/2, 1)^T$ since

$$\boldsymbol{z}_1^T A \boldsymbol{z}_2 = (1 \quad 0)\begin{pmatrix} 4 & 2 \\ 2 & 12 \end{pmatrix}\begin{pmatrix} -1/2 \\ 1 \end{pmatrix} = 0, \quad \boldsymbol{z}_1^T A \boldsymbol{z}_1 = 4 \neq 0, \quad \boldsymbol{z}_2^T A \boldsymbol{z}_2 = 13 \neq 0$$

Let us find the minimum of f generated by the point $\boldsymbol{x}_0 = (0,0)^T$ and the vectors $\boldsymbol{z}_1$, $\boldsymbol{z}_2$

Conjugate directions. Example

Example (cont.)

Starting with the $\boldsymbol{z}_1$ direction, we want to minimize

$$f(\boldsymbol{x}_0 + \alpha_1 \boldsymbol{z}_1) = f\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = f(\alpha_1, 0) = 2\alpha_1^2 + 2\alpha_1 + 3.$$

The minima $(df(\alpha_1))/d\alpha_1 = 0)$ is achieved for $\alpha_1^* = -1/2$

Proceeding now with the $\boldsymbol{z}_2$ direction, we need to minimize

$$f(\boldsymbol{x}_0 + \alpha_2 \boldsymbol{z}_2) = f\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}\right) = f(-\alpha_2/2, \alpha_2) = \frac{11}{2}\alpha_2^2 + 2\alpha_2 + 3.$$

The minima $(df(\alpha_2))/d\alpha_2 = 0)$ is achieved for $\alpha_2^* = -2/11$

So, the minimum of f is then given by

$$\boldsymbol{x}^* = \boldsymbol{x}_0 + \alpha_1^* \boldsymbol{z}_1 + \alpha_2^* \boldsymbol{z}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{2}{11}\begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{9}{22} \\ -\frac{2}{11} \end{pmatrix}$$

Conjugate gradient methods

Conjugate gradient methods generate a sequence

$$\boldsymbol{x}^k = \boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k, \quad k = 1, 2, \ldots$$

- Suposse that the directions $\boldsymbol{z}^k$ are given, and let us see first how to compute the α_k
- Define

  $$F(\alpha_k) = f(\boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k),$$

  then, the value of α_k is chosen such that

  $$\frac{dF(\alpha_k^*)}{d\alpha_k} = D_{\boldsymbol{z}^k} f(\boldsymbol{x}^{k-1} + \alpha_k^* \boldsymbol{z}^k) = (\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1} + \alpha_k^* \boldsymbol{z}^k) = (\boldsymbol{z}^k)^T \nabla f(x^k) = 0$$

Conjugate gradient methods. Quadratic functions

Assume that f is the quadratic function

$$f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T \boldsymbol{x} + \frac{1}{2} \boldsymbol{x}^T A \boldsymbol{x}$$

with A an $n \times n$ symmetric positive definite matrix. Then, from the identity

$$\boldsymbol{b} + A\boldsymbol{x}^k = \boldsymbol{b} + A\boldsymbol{x}^{k-1} + A(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$$

it follows that the gradients of f ($\nabla f(\boldsymbol{x}) = \boldsymbol{b} + A\boldsymbol{x}$) at two consecutive points ($\nabla f(\boldsymbol{x}^k) = \boldsymbol{b} + A\boldsymbol{x}^k$, $\nabla f(\boldsymbol{x}^{k-1}) = \boldsymbol{b} + A\boldsymbol{x}^{k-1}$) are related by

$$\nabla f(\boldsymbol{x}^k) = \nabla f(\boldsymbol{x}^{k-1}) + A(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$$

If $\boldsymbol{x}^k = \boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k$, we can obtain an explicit formula for α_k^* from the condition

$$\begin{aligned} 0 = \frac{dF(\alpha_k^*)}{d\alpha_k} = (\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^k) &= (\boldsymbol{z}^k)^T \left(\nabla f(\boldsymbol{x}^{k-1}) + A(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \right) \\ &= (\boldsymbol{z}^k)^T \left(\nabla f(\boldsymbol{x}^{k-1}) + \alpha_k^* A\boldsymbol{z}^k \right) \end{aligned}$$

$$\Rightarrow \quad \boxed{\alpha_k^* = -\frac{(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1})}{(\boldsymbol{z}^k)^T A\boldsymbol{z}^k}}$$

Conjugate gradient methods. Quadratic functions

Since

$$\begin{aligned} f(\boldsymbol{x}^k) &= f(\boldsymbol{x}^{k-1}) + (\boldsymbol{x}^k - \boldsymbol{x}^{k-1})^T \nabla f(\boldsymbol{x}^{k-1}) + \frac{1}{2}(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})^T A(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \\ &= f(\boldsymbol{x}^{k-1}) + \alpha_k^*(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1}) + \frac{1}{2}(\alpha_k^*)^2(\boldsymbol{z}^k)^T A\boldsymbol{z}^k \end{aligned}$$

and using the value obtained for α_k^* we get

$$f(\boldsymbol{x}^k) = f(\boldsymbol{x}^{k-1}) - \frac{(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1})}{(\boldsymbol{z}^k)^T A\boldsymbol{z}^k}(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1}) + \frac{1}{2}\left(\frac{(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1})}{(\boldsymbol{z}^k)^T A\boldsymbol{z}^k}\right)^2 (\boldsymbol{z}^k)^T A\boldsymbol{z}^k$$

From which it follows that

$$\boxed{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k-1}) = -\frac{1}{2}\frac{\left[(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1})\right]^2}{(\boldsymbol{z}^k)^T A\boldsymbol{z}^k} < 0}$$

So, assuming that the directions $\boldsymbol{z}^k$ are given, and that

$$(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1}) \neq 0 \;\Leftrightarrow\; \alpha_k^* \neq 0$$

then the **conjugate gradient method applied to the quadratic function $f(x)$ is a descent method**

Conjugate gradient methods. Choice of the directions

- We would like to do the **choice of the directions $\boldsymbol{z}^i$** in such a way that the algorithm converges fast or, even better, that terminates in a finite number of steps when applied to minimizing the quadratic function $f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}$.
- We have already seen that if the search directions $\boldsymbol{z}^k$ are mutually conjugate with respect to A, for $k = 1, ..., n$, then the point $\boldsymbol{x}^n$ will be the exact minimum of the quadratic function.
- The choice of the conjugate directions can be done in the following way:
  1. We start at a point $\boldsymbol{x}^0 \in \mathbb{R}^n$ and choose
  $$\boldsymbol{z}^1 = -\nabla f(\boldsymbol{x}^0)$$
  2. The next point, $\boldsymbol{x}^1$, is
  $$\boldsymbol{x}^1 = \boldsymbol{x}^0 + \alpha_1^* \boldsymbol{z}^1$$
  where α_1^* has been computed with the formula given in p. 24
  3. We evaluate $\nabla f(\boldsymbol{x}^1)$ and set
  $$\boldsymbol{z}^2 = -\nabla f(\boldsymbol{x}^1) + \beta_{11}\boldsymbol{z}^1,$$
  where β_{11} is such that $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ will be A-conjugate, this is
  $$(\boldsymbol{z}^1)^T A\boldsymbol{z}^2 = (\boldsymbol{z}^1)^T A\left[-\nabla f(\boldsymbol{x}^1) + \beta_{11}\boldsymbol{z}^1\right] = 0,$$
  from which it follows
  $$\beta_{11} = \frac{(\boldsymbol{z}^1)^T A\nabla f(\boldsymbol{x}^1)}{(\boldsymbol{z}^1)^T A\boldsymbol{z}^1}.$$

Conjugate gradient methods. The algorithm (cont.)

4. Once $\boldsymbol{z}^2$ is known, we determine $\boldsymbol{x}^2 = \boldsymbol{x}^1 + \alpha_2^* \boldsymbol{z}^2$, with α_2^* computed with the formula given in p. 24
5. We evaluate $\nabla f(\boldsymbol{x}^2)$ and the new direction will be
$$\boldsymbol{z}^3 = -\nabla f(\boldsymbol{x}^2) + \beta_{21}\boldsymbol{z}^1 + \beta_{22}\boldsymbol{z}^2,$$
with β_{21} and β_{22} such that $(\boldsymbol{z}^1)^T A \boldsymbol{z}^3 = (\boldsymbol{z}^2)^T A \boldsymbol{z}^3 = 0.$
6. In general, we get
$$\boldsymbol{z}^{k+1} = -\nabla f(\boldsymbol{x}^k) + \sum_{j=1}^{k} \beta_{kj}\boldsymbol{z}^j, \quad k = 0, \dots, n-1.$$

- If the function f is not quadratic, then the α_k^* can be computed using any 1-D minimization method applied to $f(\boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k)$
- If the function f is not quadratic, the matrix A at each step is an aproximation of $\nabla^2 f(\boldsymbol{x}^k)$, the Hessian matrix of f at $\boldsymbol{x}^k$, and the computation of β_{ij} is long
- We shall see later how the directions $\boldsymbol{z}^j$ can be generated more easily without the explicit use of A

Conjugate gradient methods

Theorem

Let $f(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}$ and $\boldsymbol{x}^0 \in \mathbb{R}^n$ be given, and assume that the m nonzero vectors $\boldsymbol{z}^1,\ldots,\boldsymbol{z}^m$, $\boldsymbol{z}^j \in \mathbb{R}^n$, $m \leq n$, are mutually conjugate with respect to A (symmetric and positive definite)

Starting at $\boldsymbol{x}^0$, we move to $\boldsymbol{x}^1,\ldots,\boldsymbol{x}^m$ along $\boldsymbol{z}^1,\ldots,\boldsymbol{z}^m$, respectively, such that

$$(\boldsymbol{z}^j)^T\nabla f(\boldsymbol{x}^j) = 0, \quad j = 1, \ldots, m$$

then

$$\boxed{(\boldsymbol{z}^j)^T\nabla f(\boldsymbol{x}^m) = 0, \quad j = 1, \ldots, m}$$

Corollary

If in the above theorem $m = n$, then $\nabla f(\boldsymbol{x}^n) = 0$, and x^n is the unconstrained minimum of f

Proof: Since the $\boldsymbol{z}^j$ are linearly independent, from

$$\sum_{j=1}^{n}(\boldsymbol{z}^j)^T\nabla f(\boldsymbol{x}^j) = \sum_{j=1}^{n}\nabla f(\boldsymbol{x}^j)^T\boldsymbol{z}^j = 0,$$

it follows that $\nabla f(\boldsymbol{x}^n) = 0$. $\square$

Conjugate gradient methods

Proof of the Theorem: For $j = m$ the result is obvious

Since, as we have already seen, $\nabla f(\boldsymbol{x}^k) = \nabla f(\boldsymbol{x}^{k-1}) + A(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$, it follows that the gradient of f at any two points are related by

$$\begin{aligned}\nabla f(\boldsymbol{x}^m) &= \nabla f(\boldsymbol{x}^{m-1}) + A(\boldsymbol{x}^m - \boldsymbol{x}^{m-1}) \\ &= \nabla f(\boldsymbol{x}^{m-2}) + A(\boldsymbol{x}^{m-1} - \boldsymbol{x}^{m-2}) + A(\boldsymbol{x}^m - \boldsymbol{x}^{m-1}) \\ &= \nabla f(\boldsymbol{x}^{m-2}) + A(\boldsymbol{x}^m - \boldsymbol{x}^{m-2}),\end{aligned}$$

so

$$\nabla f(\boldsymbol{x}^m) = \nabla f(\boldsymbol{x}^j) + A(\boldsymbol{x}^m - \boldsymbol{x}^j), \quad j = 1, ..., m-1. \tag{1}$$

From $\boldsymbol{x}^j = \boldsymbol{x}^{j-1} + \alpha_j^* \boldsymbol{z}^j$, for $j = 1, ..., m$, it follows that

$$\boldsymbol{x}^m = \boldsymbol{x}^{m-1} + \alpha_m^* \boldsymbol{z}^m = \boldsymbol{x}^{m-2} + \alpha_{m-1}^* \boldsymbol{z}^{m-1} + \alpha_m^* \boldsymbol{z}^m = \dots$$

so

$$\boldsymbol{x}^m - \boldsymbol{x}^j = \sum_{i=j+1}^{m} \alpha_i^* \boldsymbol{z}^i, \quad j = 0, ..., m-1$$

Conjugate gradient methods. Proof of the Theorem (cont.)

In this way, we can write

$$\nabla f(x^m) = \nabla f(x^j) + A(x^m - x^j) = \nabla f(x^j) + \sum_{i=j+1}^{m} \alpha_i^* A z^i, \quad j = 1, ..., m-1,$$

from which it follows that

$$(z^j)^T \nabla f(x^m) = (z^j)^T \nabla f(x^j) + \sum_{i=j+1}^{m} \alpha_i^* (z^j)^T A z^i = 0, \quad j = 1, ..., m-1.$$

since the first term of the right-hand side vanishes, according to the hypothesis, and the second term by the conjugacy of the z^j.

$\square$

Alternating directions methods: conjugate gradient method. Summary

Conjugate gradient methods for **quadratic functions**

$$\begin{aligned} f(\boldsymbol{x}) &= a + \boldsymbol{b}^T \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} \\ \boldsymbol{x}^k &= \boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k, \quad k = 1, 2, \ldots \end{aligned}$$

Recall that for a quadratic function f, the solution $\boldsymbol{x}^* \in \mathbb{R}^n$ is found in n steps if the search directions $\boldsymbol{z}^j$ are mutually conjugate w.r.t A

- Computation of the coefficients α_k:

$$\frac{df(\boldsymbol{x}^{k-1} + \alpha_k^* \boldsymbol{z}^k)}{d\alpha_k} = 0 \quad \Rightarrow \quad \alpha_k^* = -\frac{(\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1})}{(\boldsymbol{z}^k)^T A\boldsymbol{z}^k}$$

where

$$\nabla f(\boldsymbol{x}^{k-1}) = \boldsymbol{b} + A\boldsymbol{x}^{k-1}$$

Remark: If f is not quadratic, then the α_k^* can be computed using any 1-D minimization method applied to $f(\boldsymbol{x}^{k-1} + \alpha_k \boldsymbol{z}^k)$

Alternating directions methods: conjugate gradient method. Summary

- Computation of the search directions $\boldsymbol{z}^k$

$$
\begin{aligned}
\boldsymbol{z}^1 &= -\nabla f(\boldsymbol{x}^0) \\
\boldsymbol{z}^2 &= -\nabla f(\boldsymbol{x}^1) + \beta_{11}\boldsymbol{z}^1 \\
& \quad (\boldsymbol{z}^2)^T A \boldsymbol{z}^1 = 0 \quad \Rightarrow \beta_{11} = \frac{(\boldsymbol{z}^1)^T A \nabla f(\boldsymbol{x}^1)}{(\boldsymbol{z}^1)^T A \boldsymbol{z}^1} \\
\vdots \quad & \quad \vdots \\
\boldsymbol{z}^{k+1} &= -\nabla f(\boldsymbol{x}^k) + \sum_{j=1}^{k} \beta_{kj}\boldsymbol{z}^j
\end{aligned}
$$

where the β_{kj} for $j = 1, \ldots, k$, can be computed using the conjugate conditions

$$
(\boldsymbol{z}^{k+1})^T A \boldsymbol{z}^j = 0, \quad j = 1, 2, \ldots, k
$$

Conjugate gradient methods. General computation of the β_{ij} coefficients

- We assume that f is quadratic: $f(\boldsymbol{x}) = a + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}$
- Let
$$\boldsymbol{\gamma}^i = \nabla f(\boldsymbol{x}^i) - \nabla f(\boldsymbol{x}^{i-1}) = A(\boldsymbol{x}^i - \boldsymbol{x}^{i-1}), \quad i = 1, \ldots, n$$
- Since
$$\boldsymbol{x}^i = \boldsymbol{x}^{i-1} + \alpha_i^* \boldsymbol{z}^i \quad \Rightarrow \quad \boldsymbol{x}^i - \boldsymbol{x}^{i-1} = \alpha_i^* \boldsymbol{z}^i$$
and using that A is symmetric, it follows that
$$\boldsymbol{\gamma}^i = A(\boldsymbol{x}^i - \boldsymbol{x}^{i-1}) = \alpha_i^* A\boldsymbol{z}^i \quad \Rightarrow \quad (\boldsymbol{\gamma}^i)^T = \alpha_i^* (\boldsymbol{z}^i)^T A, \quad i = 1, \ldots, n$$
so
$$(\boldsymbol{\gamma}^i)^T \boldsymbol{z}^j = \alpha_i^* (\boldsymbol{z}^i)^T A\boldsymbol{z}^j, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n.$$
- If $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k$, $k \leq n$ are chosen to be mutually conjugate w.r.t. A, we get that for $i \neq j$
$$(\boldsymbol{\gamma}^i)^T \boldsymbol{z}^j = 0, \quad i = 1, \ldots, k, \quad j = 1, \ldots, k, \quad i \neq j$$
- We will use this last equality to obtain an expression of β_{11} independent of A

Computation of β_{11}

Recall that

$$\begin{aligned} \boldsymbol{z}^1 &= -\nabla f(\boldsymbol{x}^0) \\ \boldsymbol{z}^2 &= -\nabla f(\boldsymbol{x}^1) + \beta_{11}\boldsymbol{z}^1 \\ \gamma^1 &= \nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0) \end{aligned}$$

so, according to the last result $((\gamma^i)^T \boldsymbol{z}^j = 0, \quad i \neq j)$

$$\begin{aligned} 0 &= (\gamma^1)^T \boldsymbol{z}^2 \\ &= (\gamma^1)^T [-\nabla f(\boldsymbol{x}^1) - \beta_{11}\nabla f(\boldsymbol{x}^0)] \\ &= -(\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T (\nabla f(\boldsymbol{x}^1) + \beta_{11}\nabla f(\boldsymbol{x}^0)) \end{aligned}$$

we get

$$\beta_{11} = \frac{(\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T \nabla f(\boldsymbol{x}^1)}{(\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T (-\nabla f(\boldsymbol{x}^0))}.$$

Computation of β_{11}

On the other hand, the value of α_k^* was chosen such that

$$\frac{df(\boldsymbol{x}^{k-1} + \alpha_k^* \boldsymbol{z}^k)}{d\alpha_k} = (\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^{k-1} + \alpha_k^* \boldsymbol{z}^k) = (\boldsymbol{z}^k)^T \nabla f(\boldsymbol{x}^k) = 0$$

Recalling that

$$\boldsymbol{z}^1 \quad = \quad -\nabla f(\boldsymbol{x}^0)$$

it follows that

$$(\boldsymbol{z}^1)^T \nabla f(\boldsymbol{x}^1) = -(\nabla f(\boldsymbol{x}^0))^T \nabla f(x^1) = 0$$

so

$$\beta_{11} = \frac{(\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T \nabla f(\boldsymbol{x}^1)}{(\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T (-\nabla f(\boldsymbol{x}^0))} \quad \Rightarrow \quad \beta_{11} = \frac{(\nabla f(\boldsymbol{x}^1))^T \nabla f(\boldsymbol{x}^1)}{(\nabla f(\boldsymbol{x}^0))^T \nabla f(\boldsymbol{x}^0)}.$$

Computation of the β_{ij} coefficients

- The point $\boldsymbol{x}^2$ is reached by minimizing along the conjugate directions $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$.
- According to the last Theorem (page 28: $(\boldsymbol{z}^j)^T \nabla f(\boldsymbol{x}^m) = 0, \quad j = 1, ..., m$)
$$(\boldsymbol{z}^1)^T \nabla f(\boldsymbol{x}^2) = 0, \quad (\boldsymbol{z}^2)^T \nabla f(\boldsymbol{x}^2) = 0.$$
- Substituting $\boldsymbol{z}^1 = -\nabla f(\boldsymbol{x}^0)$ and $\boldsymbol{z}^2 = -\nabla f(\boldsymbol{x}^1) + \beta_{11}\boldsymbol{z}^1$ in these equalities, we get
$$(\nabla f(\boldsymbol{x}^0))^T \nabla f(\boldsymbol{x}^2) = 0, \quad (\nabla f(\boldsymbol{x}^1))^T \nabla f(\boldsymbol{x}^2) = 0. \tag{2}$$
- From $(\gamma^i)^T \boldsymbol{z}^j = 0$ if $i \neq j$ (see page 34) and
$$\begin{aligned}
\gamma^1 &= \nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0), \\
\gamma^2 &= \nabla f(\boldsymbol{x}^2) - \nabla f(\boldsymbol{x}^1), \\
\boldsymbol{z}^3 &= -\nabla f(\boldsymbol{x}^2) + \beta_{21}\boldsymbol{z}^1 + \beta_{22}\boldsymbol{z}^2, \\
0 &= (\gamma^1)^T \boldsymbol{z}^3 = (\nabla f(\boldsymbol{x}^1) - \nabla f(\boldsymbol{x}^0))^T(-\nabla f(\boldsymbol{x}^2) + \beta_{21}\boldsymbol{z}^1 + \beta_{22}\boldsymbol{z}^2), \\
0 &= (\gamma^2)^T \boldsymbol{z}^3 = (\nabla f(\boldsymbol{x}^2) - \nabla f(\boldsymbol{x}^1))^T(-\nabla f(\boldsymbol{x}^2) + \beta_{21}\boldsymbol{z}^1 + \beta_{22}\boldsymbol{z}^2),
\end{aligned}$$
and the equalities (2), it follows that
$$\beta_{21} = 0, \qquad \beta_{22} = \frac{(\nabla f(\boldsymbol{x}^2))^T \nabla f(\boldsymbol{x}^2)}{(\nabla f(\boldsymbol{x}^1))^T \nabla f(\boldsymbol{x}^1)}.$$

Computation of the β_{ij} coefficients

In a similar way, we can also establish that

$$\begin{aligned}\beta_{kj} &= 0, \quad \text{for} \quad k \neq j \\ \beta_{kk} &= \frac{(\nabla f(\boldsymbol{x}^k))^T \nabla f(\boldsymbol{x}^k)}{(\nabla f(\boldsymbol{x}^{k-1}))^T \nabla f(\boldsymbol{x}^{k-1})}, \quad k = 1, ..., n\end{aligned}$$

thus

$$\boxed{\boldsymbol{z}^{k+1} = -\nabla f(\boldsymbol{x}^k) + \frac{(\nabla f(\boldsymbol{x}^k))^T \nabla f(\boldsymbol{x}^k)}{(\nabla f(\boldsymbol{x}^{k-1}))^T \nabla f(\boldsymbol{x}^{k-1})} \boldsymbol{z}^k} \tag{3}$$

Remark: Note that the above equation for the direction $\boldsymbol{z}^{k+1}$ is independent of A

The conjugate gradient algorithm for a $\mathcal{C}^1$ function

1. Choose a starting point $\boldsymbol{x}^0 \in \mathbb{R}^n$.
2. Evaluate $\nabla f(\boldsymbol{x}^0)$ and set $\boldsymbol{z}^1 = -\nabla f(\boldsymbol{x}^0)$.
3. Move to $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n$
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^*_{k+1}\boldsymbol{z}^{k+1}$$
by minimizing $f(\boldsymbol{x})$ along the directions $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^n$ computed according to
$$\boldsymbol{z}^{k+1} = -\nabla f(\boldsymbol{x}^k) + \frac{(\nabla f(\boldsymbol{x}^k))^T \nabla f(\boldsymbol{x}^k)}{(\nabla f(\boldsymbol{x}^{k-1}))^T \nabla f(\boldsymbol{x}^{k-1})}\boldsymbol{z}^k$$
4. If f is quadratic, then
$$\alpha^*_{k+1} = -\frac{(\boldsymbol{z}^{k+1})^T \nabla f(\boldsymbol{x}^k)}{(\boldsymbol{z}^{k+1})^T A \boldsymbol{z}^{k+1}}$$
and the procedure finishes after the first n minimizations.
If f is not quadratic, then use any 1-D minimization procedure for the computation of α^*_{k+1}
5. After these n minimizations, restart the procedure by letting $\boldsymbol{x}^n$ and $-\nabla f(\boldsymbol{x}^n)$ be the new $\boldsymbol{x}^0$ and $\boldsymbol{z}^1$.
6. Repite the above two steps (3. and 4.) until
$$\|\nabla f(\boldsymbol{x}^k)\|^2 = (\nabla f(\boldsymbol{x}^k))^T \nabla f(\boldsymbol{x}^k) \leq \epsilon,$$
where ϵ is some predetermined small number.

The conjugate gradient algorithm. Example

Consider the quadratic function

$$f(\boldsymbol{x}) = \frac{3}{2}x^2 + \frac{1}{2}y^2 - xy - 2x,$$

so

$$\boldsymbol{a} = 0, \quad \boldsymbol{b} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \quad A = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}.$$

We take

$$\boldsymbol{x}^0 = \begin{pmatrix} -2 \\ 4 \end{pmatrix}, \quad \nabla f(\boldsymbol{x}^0) = \begin{pmatrix} -12 \\ 6 \end{pmatrix}, \quad \boldsymbol{z}^1 = -\nabla f(\boldsymbol{x}^0) = \begin{pmatrix} 12 \\ -6 \end{pmatrix}.$$

Minimizing $f(\boldsymbol{x}^0 + \alpha_1 \boldsymbol{z}^1)$ with respect to α_1 we get $\alpha_1^* = 5/17$ and

$$\boldsymbol{x}^1 = \boldsymbol{x}^0 + \alpha_1^* \boldsymbol{z}^1 = \begin{pmatrix} 26/17 \\ 38/17 \end{pmatrix}, \quad \nabla f(\boldsymbol{x}^1) = \begin{pmatrix} 6/17 \\ 12/17 \end{pmatrix}.$$

So, we have

$$\boldsymbol{z}^2 = -\nabla f(\boldsymbol{x}^1) + \frac{(\nabla f(\boldsymbol{x}^1))^T \nabla f(\boldsymbol{x}^1)}{(\nabla f(\boldsymbol{x}^0))^T \nabla f(\boldsymbol{x}^0)} \boldsymbol{z}^1 = -\begin{pmatrix} 6/17 \\ 12/17 \end{pmatrix} + \frac{(6/17)^2 + (12/17)^2}{(-12)^2 + 6^2} \begin{pmatrix} 12 \\ -6 \end{pmatrix}$$

$$= -\begin{pmatrix} 90/289 \\ 210/289 \end{pmatrix}.$$

Minimizing $f(\boldsymbol{x}^1 + \alpha_2 \boldsymbol{z}^2)$ with respect to α_2 we get $\alpha_2^* = 17/10$. Consequently $\boldsymbol{x}^2 = \boldsymbol{x}^1 + \alpha_2^* \boldsymbol{z}^2 = (1, 1)^T$, which is the global minimum of the quadratic function f.

The conjugate gradient method. Exercises

Exercise 5. To be delivered before 2-XI-2021 as: `Ex05-YourSurname.pdf`

Solve the linear system

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

using the conjugate-gradient method.

Exercise 6. To be delivered before 2-XI-2021 as: `Ex06-YourSurname.pdf`

Consider the conjugate gradient method applied to the minimization of

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

where A is a positive definite and symmetric matrix.
Show that the iterate $\boldsymbol{x}^k$ minimizes f over

$$\boldsymbol{x}^0 + < \boldsymbol{v}^0, A\boldsymbol{v}^0, ..., A^{k-1}\boldsymbol{v}^0 >$$

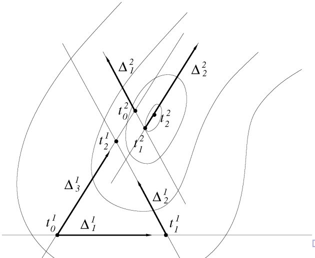
where $\boldsymbol{v}^0 = \nabla f(\boldsymbol{x}^0)$, and $< v^0, A\boldsymbol{v}^0, ..., A^{k-1}\boldsymbol{v}^0 >$ is the subspace generated by $v^0, A\boldsymbol{v}^0, ..., A^{k-1}\boldsymbol{v}^0$

Powell's method (for continuous functions)

- We start presenting Powell's method as an empirical technique
- The method does not require the computation of derivatives and, from now on, we will not assume that $f(\boldsymbol{x})$ is a quadratic function
- The basic version of the method is as follows:
  1. Each stage the procedure consists of $\boldsymbol{n+1}$ successive 1-dimensional line searches
  2. The first n searches are done along n linearly independent directions
  3. The $(n+1)$th search is done along the direction connecting:
     - the obtained best point (obtained at the end of the n preceeding 1-dimensional line searches)
     - with the starting point of that stage
  4. After these $n+1$ searches, one of the first n directions is replaced by the $(n+1)$–th direction, and a new stage begins

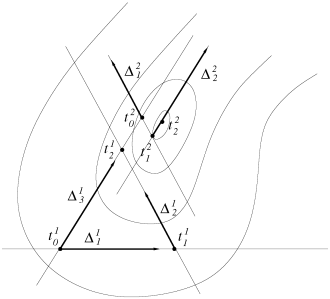
The k–th stage of Powell's method

1. Let $\boldsymbol{x}_B^{k-1} = \boldsymbol{t}_0^k \in \mathbb{R}^n$ be the starting point of the k–th stage and $\Delta_1^k, \ldots, \Delta_n^k$, n linearly independent directions. *(for $n = 2$, $k = 1$, start with $\boldsymbol{t}_0^1$, Δ_1^1, Δ_2^1)*
2. Determine θ_j^*, for $j = 1, \ldots, n$ *(for $n = 2$, k=1, determine θ_1^* and θ_2^*)* such that
$$f(\boldsymbol{t}_{j-1}^k + \theta_j^* \Delta_j^k) = \min_{\theta_j} f(\boldsymbol{t}_{j-1}^k + \theta_j \Delta_j^k),$$
and define
$$\boldsymbol{t}_j^k = \boldsymbol{t}_{j-1}^k + \theta_j^* \Delta_j^k, \quad j = 1, \ldots, n.$$
(for $n = 2$, $k = 1$: define $\boldsymbol{t}_1^1 = \boldsymbol{t}_0^1 + \theta_1^* \Delta_1^1, \quad \boldsymbol{t}_2^1 = \boldsymbol{t}_1^1 + \theta_2^* \Delta_2^1$)
3. The new search directions are
$$\Delta_j^{k+1} = \Delta_{j+1}^k, \quad j = 1, \ldots, n-1, \quad \Delta_n^{k+1} = \Delta_{n+1}^k = \boldsymbol{t}_n^k - \boldsymbol{t}_0^k.$$
(for $n = 2$, $k = 1$, the new directions are $\Delta_1^2 = \Delta_2^1$, $\Delta_2^2 = \Delta_3^1 = \boldsymbol{t}_2^1 - \boldsymbol{t}_0^1$, and the starting point stage is t_2^1)



The k-th stage of Powell's method

4. Find θ^*_{n+1} such that
$$f(\boldsymbol{t}_n^k + \theta^*_{n+1}(\boldsymbol{t}_n^k - \boldsymbol{t}_0^k)) = \min_{\theta_{n+1}} f(\boldsymbol{t}_n^k + \theta_{n+1}(\boldsymbol{t}_n^k - \boldsymbol{t}_0^k)),$$
(for $n = 2$, $k = 1$: find θ^*_3 s.t. $f(\boldsymbol{t}_2^1 + \theta^*_3 \Delta_2^2) = \min_{\theta_3} f(\boldsymbol{t}_2^1 + \theta_3 \Delta_2^2)$: $t_0^2 = \boldsymbol{t}_2^1 + \theta_3 \Delta_2^2$)



5. Take as new initial point
$$\boldsymbol{x}_B^k = \boldsymbol{t}_0^{k+1} = \boldsymbol{t}_n^k + \theta^*_{n+1}(\boldsymbol{t}_n^k - \boldsymbol{t}_0^k).$$
(for $n = 2$, $k = 1$, take as initial point $\boldsymbol{x}_B^1 = t_0^2 = \boldsymbol{t}_2^1 + \theta^*_3 \Delta_2^2$ and explore along the directions Δ_1^2 and Δ_2^2)
6. If $\|\boldsymbol{x}_B^{k-1} - \boldsymbol{x}_B^k\| < \epsilon$ ($\epsilon > 0$ fixed) stop, otherwise proceed to stage $k+1$.

Powell's method. Example 1 (first step)

Let

$$f(x,y) = \frac{3}{2}x^2 + \frac{1}{2}y^2 - xy - 2x,$$

which has a minimum at $(1,\ 1)$.

1. We start with

$$\boldsymbol{x}_B^0 = \boldsymbol{t}_0^1 = \begin{pmatrix} -2 \\ 4 \end{pmatrix}, \quad \Delta_1^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \Delta_1^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

2. The first minimization is in the Δ_1^1 direction

$$\min_{\theta_1} f(\boldsymbol{t}_0^1 + \theta_1 \Delta_1^1) = \min_{\theta_1} \left\{ \frac{3}{2}(-2+\theta_1)^2 + \frac{1}{2}4^2 - (-2+\theta_1)4 - 2(-2+\theta_1) \right\}$$

$$\Rightarrow \quad \theta_1^* = 4, \quad \Rightarrow \quad t_1^1 = (2,\ 4)^T$$

3. Now we minimize in the Δ_2^1 direction

$$\min_{\theta_2} f(\boldsymbol{t}_1^1 + \theta_2 \Delta_2^1) = \min_{\theta_2} \left\{ \frac{3}{2}2^2 + \frac{1}{2}(4+\theta_2)^2 - 2(4+\theta_2) - 4 \right\}$$

$$\Rightarrow \quad \theta_2^* = -2, \quad \Rightarrow \quad t_2^1 = (2,\ 2)^T$$

4. Consequently, the new direction is

$$\Delta_3^1 = t_2^1 - t_0^1 = \begin{pmatrix} 2-(-2) \\ 2-4 \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

Powell's method. Example 1 (first step)

5. Next we minimize along the new direction Δ_3^1

$$\min_{\theta_3} f(\boldsymbol{t}_2^1 + \theta_3 \Delta_3^1) =$$

$$= \min_{\theta_3} \left\{ \frac{3}{2}(2+4\theta_3)^2 + \frac{1}{2}(2-2\theta_3)^2 - (2+4\theta_3)(2-2\theta_3) - 2(2+4\theta_3) \right\}$$

$$\Rightarrow \quad \theta_3^* = -2/17, \quad \Rightarrow \quad x_B^1 = t_0^2 = \begin{pmatrix} 2-8/17 \\ 2+4/17 \end{pmatrix} = \begin{pmatrix} 26/17 \\ 38/17 \end{pmatrix}$$

This concludes the first iteration of the algorithm. The first two search directions of the second iteration are

$$\Delta_1^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \Delta_2^2 = \begin{pmatrix} 4 \\ -2 \end{pmatrix}.$$

Powell's method. Example 1 (second step)

6. The first minimization is in the Δ_1^2 direction

$$\min_{\theta_1} f(\boldsymbol{t}_0^2 + \theta_1 \Delta_1^2) = \min_{\theta_1} \left\{ \frac{3}{2}\left(\frac{26}{17}\right)^2 + \frac{1}{2}\left(\frac{38}{17} + \theta_1\right)^2 - \frac{26}{17}\left(\frac{38}{17} + \theta_1\right) - \frac{52}{17} \right\}$$

$$\Rightarrow \quad \theta_1^* = -12/17, \quad \Rightarrow \quad \boldsymbol{t}_1^2 = (26/17,\ 26/17)^T.$$

7. The second minimization is in the Δ_2^2 direction

$$\min_{\theta_2} f(\boldsymbol{t}_1^2 + \theta_2 \Delta_2^2) =$$

$$= \min_{\theta_2} \left\{ \frac{3}{2}\left(\frac{26}{17} + 4\theta_2\right)^2 + \frac{1}{2}\left(\frac{26}{17} - 2\theta_2\right)^2 - \left(\frac{26}{17} + 4\theta_2\right)\left(\frac{26}{17} - 2\theta_2\right) - 2\left(\frac{26}{17} + 4\theta_2\right) \right\}$$

$$\Rightarrow \quad \theta_2^* = -18/289, \quad \Rightarrow \quad t_2^2 = (370/289,\ 478/289)^T.$$

8. The new direction is

$$\Delta_3^2 = t_2^2 - t_0^2 = \begin{pmatrix} -72/289 \\ -168/289 \end{pmatrix}.$$

9. Finally, when we compute $\min_{\theta_3} f(\boldsymbol{t}_2^2 + \theta_3 \Delta_3^2) = \ldots$ we get

$$\theta_3^* = 9/8, \quad x_B^2 = (1,\ 1)^T.$$

That is, the exact minimum of the quadratic function is found in two iterations

Powell's method. Example 2

In the above example the directions $\Delta_{1,2}^{k}$ $(k = 1, 2)$ were linearly independent. This condition is important, as is shown in the next example.

Let

$$f(x, y, z) = (x - y + z)^2 + (-x + y + z)^2 + (x + y - z)^2,$$

that has a minimum at $(x^*, y^*, z^*) = (0,\ 0,\ 0)$. Start Powell's method with

$$x_B^0 = \begin{pmatrix} 1/2 \\ 1 \\ 1/2 \end{pmatrix}, \quad \Delta_1^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \Delta_2^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \Delta_3^1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The results of the first three steps are

j	Δ_j^1	t_j^1	$f(t_j^1)$
0	–	$(1/2,\ 1,\ 1/2)$	2
1	Δ_1^1	$(1/2,\ 1,\ 1/2)$	2
2	Δ_2^1	$(1/2,\ 1/3,\ 1/2)$	$2/3$
3	Δ_3^1	$(1/2,\ 1/3,\ 5/18)$	$42/81$

The new direction is

$$\boldsymbol{t}_3^1 - \boldsymbol{t}_0^1 = \begin{pmatrix} 1/2 \\ 1/3 \\ 5/18 \end{pmatrix} - \begin{pmatrix} 1/2 \\ 1 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 0 \\ -2/3 \\ -2/9 \end{pmatrix}$$

Powell's method. Example 2 (cont.)

The new search directions are

$$\Delta_1^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \Delta_2^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \Delta_3^2 = \begin{pmatrix} 0 \\ -2/3 \\ -2/9 \end{pmatrix}.$$

Thus, the first component of all forthcoming points reached will remain equal to $1/2$, and the true optimum at $(x^*, y^*, z^*) = (0,\ 0,\ 0)$ can never be reached.

Powell's method for quadratic functions

Let us show how the properties about conjugate directions can be used to prove termination of Powell's method in a finite number of steps for quadratic functions.

Assume that:

- The function f is quadratic and A is a symmetric positive define matrix.
- The initial point is $\boldsymbol{x}_B^0 \in \mathbb{R}^n$.
- The initial directions $\Delta_1^1, \ldots, \Delta_n^1$ are linearly independent.

After the steps of the first stage, we have:

- The $n+1$ points: $t_0^1, \ldots, t_n^1$.
- A new direction $\Delta_n^2 = \boldsymbol{z}^1 = \boldsymbol{t}_n^1 - \boldsymbol{t}_0^1$. We assume that $\boldsymbol{t}_n^1 \neq \boldsymbol{t}_0^1$.
- A new starting point $\boldsymbol{t}_0^2 = \boldsymbol{t}_n^1 + \theta_{n+1}^* \Delta_n^2 = \boldsymbol{x}_B^1$.
- The new starting point $\boldsymbol{x}_B^1 = \boldsymbol{t}_0^2$, is a minimum of f in the Δ_n^2 direction.

Powell's method for quadratic functions

- Because of the properties of the conjugate directions (see Theorem in page 18), the direction $\boldsymbol{z}^2 = \boldsymbol{t}_n^2 - \boldsymbol{t}_0^2$ is conjugate to $\boldsymbol{z}^1$ with respect to A.
- After k steps of the procedure, we have generated k non-zero directions $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k$ mutually conjugate w.r.t. A.
- If the directions $\Delta_1^k, \ldots, \Delta_{n-k}^k$, $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k$ are linearly independent, then $\boldsymbol{z}^{k+1} = \boldsymbol{t}_n^{k+1} - t_0^{k+1}$ will be conjugate to $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k$.
- After completing n stages all the search directions are mutually conjugate w.r.t. A and the minimum of f over $\mathbb{R}^n$ has been reached.

  Recall that for a quadratic function, if a point in $\mathbb{R}^n$ is optimal in n mutually conjugate t directions, then it must be the global optimum of the function (see Theorem in page 20).

Avoiding linearly dependent search directions

We can modify Powell's method to avoid linearly dependent search directions.

The new method does not possess the quadratic termination property, but has a satisfactory performance.

Let

- $\boldsymbol{x}_B^{k-1} = \boldsymbol{t}_0^k$ be the starting point of the k–th stage.
- $\Delta_1^k, \ldots, \Delta_n^k$, n linearly independent directions.

Then

- Find $\boldsymbol{t}_j^k$ for $j = 1, ..., n$, the minima of f along the directions $\Delta_1^k, \ldots, \Delta_n^k$.
- Set $\Delta_{n+1}^k = \boldsymbol{t}_n^k - \boldsymbol{t}_0^k$.
  - If $\|\boldsymbol{t}_n^k - \boldsymbol{t}_0^k\| < \epsilon$, stop.
  - Otherwise, find α_{n+1}^* such that

$$f(\boldsymbol{t}_0^k + \alpha_{n+1}^* \Delta_{n+1}^k) = \min_{\alpha_{n+1}} f(\boldsymbol{t}_0^k + \alpha_{n+1} \Delta_{n+1}^k),$$

and let $\boldsymbol{t}_0^{k+1} = \boldsymbol{x}_B^k = \boldsymbol{t}_0^k + \alpha_{n+1}^*$.

Avoiding linearly dependent search directions

- If $\|\boldsymbol{x}_B^k - \boldsymbol{x}_B^{k-1}\| < \epsilon$, stop (convergence).
- Otherwise find the index m such that

$$f(\boldsymbol{t}_{m-1}^k) - f(\boldsymbol{t}_m^k) = \max_{j=1,\dots,n} \{f(\boldsymbol{t}_{j-1}^k) - f(t_j^k)\},$$

(largest function decreese).

  - If
  $$|\alpha_{n+1}^*| < \left(\frac{f(\boldsymbol{t}_0^k) - f(\boldsymbol{t}_0^{k+1})}{f(\boldsymbol{t}_{m+1}^k) - f(\boldsymbol{t}_m^k)} \right)^{1/2}, \tag{4}$$
  set $\Delta_j^{k-1} = \Delta_j^k$, $j = 1, \dots, n$.
  In other words, the search directions of the $(k+1)$–th stage are the same as in the k–th stage.

  - If (4) does not hold, set
  $$\begin{aligned} \Delta_j^{k-1} &= \Delta_j^k, \quad j = 1, \dots, m-1, \\ \Delta_j^{k-1} &= \Delta_{j+1}^k, \quad j = m, \dots, n, \end{aligned}$$
  and proceed to stage $k+1$.

Avoiding linearly dependent search directions. Example 2 (cont.)

Consider again the problem of Example 2.

$$f(x,y,z) = (x-y+z)^2 + (-x+y+z)^2 + (x+y-z)^2.$$

The first steps are the same as before. We can see that the largest function decrease is obtained by going from t_1^1 to t_2^1, hence $m = 2$.
We have

$$\Delta_4^1 = (0,\ -2/3,\ -2/9)^T.$$

We find that $\alpha_4^* = 9/8$ minimizes

$$f(1/2,\ 1-(2/3)\alpha_4,\ 1/2-(2/9)\alpha_4) \quad \Rightarrow$$

$$\boldsymbol{t}_0^2 = (1/2,\ 1,\ 1/2)^T + (9/8)(0,\ -2/3,\ -2/9)^T = (1/2,\ 1/4,\ 1/4)^T \quad \Rightarrow f(\boldsymbol{t}_0^2) = 1/2.$$

Now

$$\left(\frac{f(\boldsymbol{t}_0^1) - f(\boldsymbol{t}_0^2)}{f(\boldsymbol{t}_1^1) - f(\boldsymbol{t}_2^1)}\right)^{1/2} = \left(\frac{2-1/2}{2-2/3}\right)^{1/2} = (9/8)^{1/2}.$$

Since $\alpha_4^* > (9/8)^{1/2}$, we see that (4) does not hold. Accordingly, the new directions will be the independents vectors

$$\begin{aligned} \Delta_1^2 &= (1,\ 0,\ 0)^T, \\ \Delta_2^2 &= (0,\ 0,\ 1)^T, \\ \Delta_3^2 &= (0,\ -2/3,\ -2/9)^T. \end{aligned}$$