PRO 5970 Métodos de Otimização Não Linear

Celma de Oliveira Ribeiro

2023

Departamento de Engenharia de Produção
Universidade de São Paulo

Newton Search

Advantages:

- Excellent performance of Newton search close to the optimum
- Less sensitive to numerical errors than steepest descent search

Disadvantage:

- Very sensitive to starting point x_0
- Can fail to converge when starting relatively far from a local optimum!
- Hessian matrix needed at each iteration, as well as solution of a linear system - Very burdensome task, especially for large-scale systems!

Need to mitigate these deficiencies!

Conjugate directions

- Conjugate direction methods: intermediate between the method of steepest descent and Newton's method.
- They typically perform better than the method of steepest descent, but not as well as Newton's method.

Properties of conjugate direction methods

- Solve quadratics of n variables in n steps
- Conjugate gradient algorithm requires no Hessian matrix evaluations
- No matrix inversion and no storage of $n \times n$ matrix are require

Search directions

The optimization methods considered usually find, at iteration k, a direction d_k, such that

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

For a given function f

- Steepest descent

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

- Newton

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

For quadratic function $f(x) = \frac{1}{2} x^t A x - b^t x$,

$$\nabla f(x_k) = Ax - b$$

and

$$H(x_k) = A$$

For quadratic functions some nice results can be easily found

Consider a set of directions $\{d_k\}$and the line search approach

Proposition 1 - Quadratic functions

Consider a quadratic function $f(x) = \frac{1}{2}x^{'}Ax - b^{'}x$, with A symmetric positive definite, $d_k \in \mathbb{R}^n$, $d_k \neq 0$ and $\phi(\lambda) = f(x_k + \lambda d_k)$ (line search)

The optimal solution of $\min_{\lambda \in \mathbb{R}} \phi(\lambda)$ is

$$\lambda^* = -\frac{\nabla f(x_k)^t d_k}{d_k^t A d_k}$$

This result holds for any direction $d_k \neq 0$!!!!

$\boxed{\text{Numerical methods are not necessary}}$

Exercice Entrega - aula

Verify that the proposition is valid for

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Proposition 2 - Quadratic functions

For the steepest descent, the solution for function f (quadratic) is is

$$\lambda^* = \frac{\nabla f(x_k)^t \nabla f(x_k)}{\nabla f(x_k)^t A \nabla f(x_k)}$$

See that this is a special case of the previous proposition!

Interpretation of the properties of conjugate directions.

Consider $\frac{1}{2}x^t Ax - b^t x$

If the matrix A is diagonal, the contours of the function $\phi(x)$ are ellipses whose axes are aligned with the coordinate directions, as illustrated below.We can find the minimizer of this function by performing one-dimensional minimizations along the coordinate directions

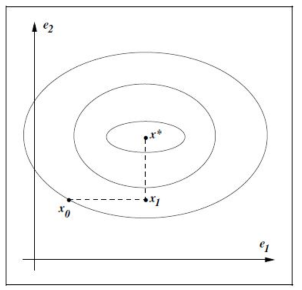


Figure 1: Successive minimization along coordinate axes does not find the solution in n iterations, for a general convex quadratic.

Interpretation of the properties of conjugate directions.

When A is not diagonal, its contours are still elliptical, but they are usually no longer aligned with the coordinate directions. The strategy of successive minimization along these directions in turn no longer leads to the solution in n iterations (or even in a finite number of iterations).

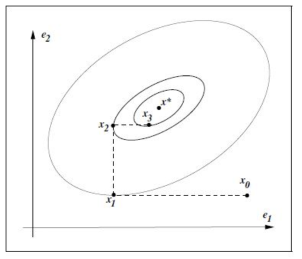


Figure 2: successive minimizations along the coordinate directions find the minimizer of a quadratic with a diagonal Hessian in n iterations.

Conjugacy

A set of non zero vectors $\{p_0, p_1, \dots, p_k\}$ is said to be *conjugate* with respect to the symmetric positive definite matrix A if

$$p_i^t A p_j = 0 \quad \forall i \neq j$$

Example 1

Let

$$A = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

i Show that A is positive definite

ii Considering $d_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}'$ construct a set of A conjugate vectors, d_0, d_1, d_2

- **Positive definite** $det(A_1) = 3 > 0 \qquad det(A_2) = 12 > 0 \qquad det(A_3) = 20 > 0$

Conjugate Direction Methods

- **Positive definite** $det(A_1) = 3 > 0 \quad det(A_2) = 12 > 0 \quad det(A_3) = 20 > 0$
- $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} = 0 \Rightarrow 3d_{11} + d_{31} = 0$

  Let $d_1 = \begin{bmatrix} 1 & 0 & -3 \end{bmatrix}'$

- **Positive definite** $det(A_1) = 3 > 0 \qquad det(A_2) = 12 > 0 \qquad det(A_3) = 20 > 0$
- $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} = 0 \Rightarrow 3d_{11} + d_{31} = 0$

  Let $d_1 = \begin{bmatrix} 1 & 0 & -3 \end{bmatrix}'$
- d_2 such that

$$d_0^{'} A d_2 = 0 \Rightarrow 3d_{12} + d_{32} = 0$$

$$d_1^{'} A d_2 = 0 \Rightarrow -6d_{22} - 8d_{32} = 0$$

  $d_2 = \begin{bmatrix} 1 & 4 & -3 \end{bmatrix}'$

Conjugate direction methods

- The search direction d_k in iteration k is conjugate to previous ones $(d_1, d_2, \dots, d_{k-1})$.
- Between the method of steepest descent and Newton's method.
- Implementation requires no Hessian matrix evaluations, no inversions or storage.
- It performs better than *Steepest Descent*, but not as well as *Newton's method*.

Example 2

Entregar

Consider the quadratic function

$$f(x) = -12x_2 + 4x_1^2 + 4x_2^2 + 4x_1x_2$$

Hessian:

$$\begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}$$

Let $d_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}'$

- Find a conjugate direction

The Conjugate Direction Algorithm - Quadratic functions

Consider $\frac{1}{2}x^t Ax - b^t x$, with with A symmetric positive definite, $x \in \mathbb{R}^n$

Because A is symmetric positive definite the function has a global minimizer that can be found by solving $Ax = b$

Basic Conjugate Direction Algorithm.

Given a starting point $x(0)$ and and n A-conjugate directions $(d_1, d_2, \ldots, d_{n-1})$.; for $k \geq 0$

$$g_k = \nabla f(x_k) = Ax_k - b$$
$$\alpha_k = -\frac{g_k' d_k}{d_k' A d_k}$$
$$x_{k+1} = x_k + \alpha_k d_k$$

Theorem For any starting point x_0 , the basic conjugate direction algorithm converges to the unique x^* (that solves $Ax = b$) in n steps; that is, $x_n = x^*$.

Example

Find the minimizer of

$$f(x_1, x_2) = \frac{1}{2}x^t \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} x - \begin{bmatrix} -1 \\ 1 \end{bmatrix} x$$

using the conjugate direction method with the initial point $x_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^t$ and A-conjugate direction $d_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^t$ and $d_1 = \begin{bmatrix} -\frac{3}{8} & \frac{3}{4} \end{bmatrix}^t$

First step

First step

In this case

$$g_0 = Ax_0 - b = \left[\begin{array}{cc} 1 & -1 \end{array} \right]^t$$

$$\alpha_0 = -\frac{g_0^t d_k}{d_0' A d_0} = -\frac{1}{4}$$

$$x_1 = x_0 + \alpha_0 d_0 = \left[\begin{array}{cc} -\frac{1}{4} & 0 \end{array} \right]^t$$

Second step

First step

In this case

$$g_0 = Ax_0 - b = \begin{bmatrix} 1 & -1 \end{bmatrix}^t$$

$$\alpha_0 = -\frac{g_0^t d_k}{d_0^{'} A d_0} = -\frac{1}{4}$$

$$x_1 = x_0 + \alpha_0 d_0 = \begin{bmatrix} -\frac{1}{4} & 0 \end{bmatrix}^t$$

Second step

$$g_1 = Ax_1 - b = \begin{bmatrix} 0 & -\frac{3}{2} \end{bmatrix}^t$$

$$\alpha_1 = -\frac{g_1^t d_1}{d_1^{'} A d_1} = 2$$

$$x_2 = x_1 + \alpha_1 d_1 = \begin{bmatrix} -1 & \frac{3}{2} \end{bmatrix}^t$$

Because f is quadratic and A is positive definite, $x_2 = x^*$

First step

In this case

$$g_0 = Ax_0 - b = \begin{bmatrix} 1 & -1 \end{bmatrix}^t$$

$$\alpha_0 = -\frac{g_0^t d_k}{d_0' A d_0} = -\frac{1}{4}$$

$$x_1 = x_0 + \alpha_0 d_0 = \begin{bmatrix} -\frac{1}{4} & 0 \end{bmatrix}^t$$

Second step

$$g_1 = Ax_1 - b = \begin{bmatrix} 0 & -\frac{3}{2} \end{bmatrix}^t$$

$$\alpha_1 = -\frac{g_1^t d_1}{d_1' A d_1} = 2$$

$$x_2 = x_1 + \alpha_1 d_1 = \begin{bmatrix} -1 & \frac{3}{2} \end{bmatrix}^t$$

Because f is quadratic and A is positive definite, $x_2 = x^*$

Exercice - Entrega

Repeat the previous exercise for

a. Two alternative initial points

b. The initial direction $\begin{bmatrix} 0 & 1 \end{bmatrix}^t$ (in this case find the conjugate direction and consider the two initial points proposed in a.)

A special case of the Conjugate Direction Method

Ref Luenberger

- In the conjugate gradient method the directions are not specified beforehand, but rather are determined sequentially at each step of the iteration.
- At step k one evaluates the current negative gradient vector and adds to it a linear combination of the previous direction vectors to obtain a new conjugate direction vector along which to move.

Advantages

- is the especially simple formula that is used to determine the new direction vector. This simplicity makes the method only slightly more complicated than steepest descent.
- because the directions are based on the gradients, the process makes good uniform progress toward the solution at every step. This is in contrast to the situation for arbitrary sequences of conjugate directions in which progress may be slight until the final few steps
- Although for the pure quadratic problem uniform progress is of no great importance, it is important for generalizations to nonquadratic problems.

A special case of the Conjugate Direction Method

Basic Properties

- The conjugate direction method is very effective. However, we need to specify the conjugate directions.
- The conjugate gradient algorithm does not use pre specified conjugate directions, but instead computes the directions as the algorithm proceeds.
- At each stage , the direction is calculated as a linear combination of the previous direction and the current gradient, in such as way that all the directions are mutually -conjugate.
- To generate the conjugate vectors, a new d_k is obtained by using only the previous vector d_{k-1} The new vector is automatically conjugate to all the previous elements

- The conjugate gradient method is an iterative method for solving a linear system of equations $Ax = b$ where A is an $n \times n$ symmetric positive definite matrix.
- This problem can be stated equivalently as the following minimization problem:

$$\min \phi(x) = \frac{1}{2}x^{'}Ax - b^{'}x$$

- One of the remarkable properties of the conjugate direction methods is its ability to generate, in a very economical fashion, a set of vectors with the property of **conjugacy**

The algorithm (quadratic functions)

Given x_0; set $g_0 = Ax_0 - b$; $d_0 = -g_0$; $k \leftarrow 0$;

While $g_k \neq 0$

$$\alpha_k \leftarrow -\frac{g_k' d_k}{d_k' A d_k}$$

$$x_{k+1} \leftarrow x_k + \alpha_k d_k$$

$g_{k+1} \leftarrow Ax_{k+1} - b$ (This is the gradient at x_{k+1})

$$\beta_{k+1} \leftarrow \frac{g_{k+1}' A d_k}{d_k' A d_k}$$

$$d_{k+1} \leftarrow -g_{k+1} + \beta_{k+1} d_k$$

end (while)

The sequence $\{x_k\}$converges to x^* in at most n steps.

It is possible to verify that the algorithm is a conjugate direction algorithm,

Example 4

Entregar

Consider the quadratic function $f(x) = \frac{1}{2}x^{'}Ax - b^{'}x$ with

$$A = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}, b = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$$

Find the minimizer using the conjugate gradient algorithm. Starting point $x_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^{'}$

The Conjugate Gradient for Non quadratic Problems

- When applied to nonquadratic problems, conjugate gradient methods will not usually terminate within n steps. It is possible therefore simply to continue finding new directions according to the algorithm and terminate only when some termina- tion criterion is met.
- The algorithm can be extended to general nonlinear functions by interpreting $f(x) = \frac{1}{2}x^{'}Ax - b^{'}x$ as a second-order Taylor series approximation of the objective function.
- For a quadratic function the Hessian is constant. However, for a general nonlinear function the Hessian is a matrix that has to be reevaluated at each iteration
- Observe that A appears only in the computation of the scalars α_k and β_k .
- Two simple changes in the preceding algorithm

  The α_k is obtained through line search

  Vector g_k is the gradient of the non linear function f

Fletcher-Reeves Algorithm

Given x_0;

Set $\nabla f_0 = \nabla f(x_0)$;$f_0 = f(x_0)$ $d_0 = -\nabla f_0$; $k \leftarrow 0$;

While $\nabla f_k \neq 0$

Compute α_k, the optimal solution of $\min_{\alpha \geq 0} f(x_j + \alpha d_j)$
$x_{k+1} \leftarrow x_k + \alpha_k d_k$
Evaluate $d_{k+1} = -\nabla f_{k+1}$
$\beta_{k+1} \leftarrow \frac{\nabla f_{k+1}' \nabla f_{k+1}}{\nabla f_k' \nabla f_k}$
$d_{k+1} \leftarrow -\nabla f_{k+1} + \beta_{k+1} d_k$

end (while)

Usually a restarting procedure is included and after n steps the process is reestarted with a pure gradient step. Thus the following step is considered

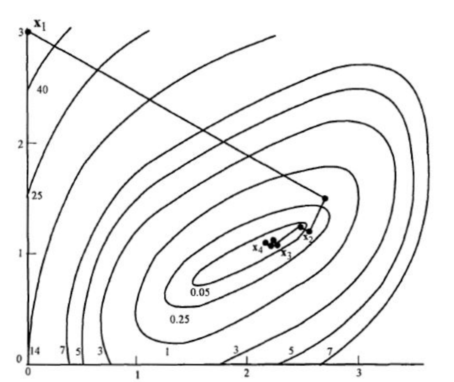
Replace x_0 by x_n and go back to Step 1.

Example 8.8.7 Bazaraa. A few differences... Notation His λ = our α. His α is our β

$min\,(x_1 - 2)^4 + (x_1 - 2x_2)^2$

Table 8.14 Summary of Computations for the Method of Fletcher and Reeves

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\lVert\nabla f(\mathbf{y}_j)\rVert$	α_{j-1}	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	(–44.00, 24.00)	50.12	—	(44.00, –24.00)	0.062	(2.70, 1.51)
		2	(2.70, 1.51) 0.34	(0.73, 1.28)	1.47	0.0009	(–0.69, –1.30)	0.23	(2.54, 1.21)
2	(2.54, 1.21) 0.10	1	(2.54, 1.21) 0.10	(0.87, –0.48)	0.99	—	(–0.87, 0.48)	0.11	(2.44, 1.26)
		2	(2.44, 1.26) 0.04	(0.18, 0.32)	0.37	0.14	(–0.30, –0.25)	0.63	(2.25, 1.10)
3	(2.25, 1.10) 0.008	1	(2.25, 1.10) 0.008	(0.16, –0.20)	0.32	—	(–0.16, 0.20)	0.10	(2.23, 1.12)
		2	(2.23, 1.12) 0.003	(0.03, 0.04)	0.05	0.04	(–0.036, –0.032)	1.02	(2.19, 1.09)
4	(2.19, 1.09) 0.0017	1	(2.19, 1.09) 0.0017	(0.05, –0.04)	0.06	—	(–0.05, 0.04)	0.11	(2.185, 1.094)
		2	(2.185, 1.094) 0.0012	(0.002, 0.01)	0.02				


x1
3
2
1
0
0
1
2
3
40
25
14
7
5
3
1
3
5
7
0.05
0.25
x2
x3
x4

Comments on Non Linear Conjugate Gradient Methods

Advantages

- use relatively little memory for large-scale problems
- require no numerical linear algebra, so each step is quite fast.

Disadvantages

- Typically converge much more slowly than Newton or quasi-Newton methods.
- steps are typically poorly scaled for length, so the line search algorithm may require more iterations each time to find an acceptable step.

Fletcher and Reeves can perform better if it is periodically restarted along the steepest descent direction

Interpretation of the properties of conjugate directions.

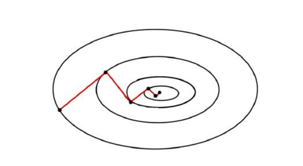


Figure 3: The contour plot of a function, with the steps of the steepest descent method in red

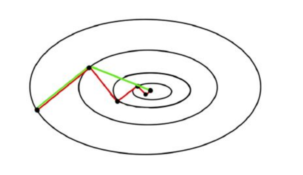


Figure 4: The contour plot of a function, with the steps of the steepest descent method in red and of the conjugate gradient method in green

Exercise - week (17/07, 23:59)

Consider the quadratic form $f(x) = c^t x + \frac{1}{2} x^t H x$ with H a symmetric $n \times n$ matrix. In many applications, it is desirable to obtain separability in the variables by eliminating the cross- product terms.This could be done by rotating the axes as follows. Let D be an $n \times n$ matrix whose columns $d_1, d_2, ...d_n$ are H conjugate. Let $x = Dy$.

a) Give an example of quadratic problem ($n \geq 2$, non-trivial) and build the H-conjugate directions

b) For the example, verify that with the rotation, the quadratic form is equivalent to $\sum_{j=1}^{n} \alpha_j y_j + \frac{1}{2} \sum_{j=1}^{n} \beta_j y_j^2$, where $\beta_j = d_j^t H d_j$ and $(\alpha_1 \alpha_2 ... \alpha_n) = c^t D$ for $j = 1, 2...n$

Translating and rotating the axes could be accomplished by the transformation $x = Dy + z$, where z is any vector satisfying $Hz + c = 0$, that is, $\nabla f(x) = 0$. In this case it can be shown that the quadratic form is equivalent to
$= c^t x + \frac{1}{2} z^t H z + \frac{1}{2} \sum_{j=1}^{n} \beta_j y_j^2$

c) Use this result to draw accurate contours of the quadratic form $2x_1 - 4x_2 + x_1^2 + 2X_1 x_2 + 3x_2^2$

Exercice weekly Consider the quadratic form
$\frac{3}{2}x_1^2 + 2x_2^2 + \frac{3}{2}x_3^2 + x_1x_3 + 2x_2x_3 - 3x_1 - x_3$

Find the minimizer using conjugate gradient algorithm with starting point
$x_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}'$