Unconstrained optimization
Conjugate gradient methods

2 октября 2017 г.

Conjugate gradient methods

Methods of conjugate directions is based on properties of vectors which conjugate with respect to a positive definite square matrix.
The difference in the methods of conjugate directions is in construction of a system of conjugate vectors for determining of the descent direction.
There are several algorithms.

Methods of conjugate directions compared with gradient methods have higher convergence rate.

Minimum of a positive definite quadratic function of n variables is for n steps.

Two non-zero vector $g_1, g_2 \in \mathbb{R}^n$ are called conjugate with respect to a positive definite matrix A size of $(n \times n)$ or A -conjugate if the scalar product $\langle Ag_1, g_2 \rangle = 0$.

Lemma 1.

If vectors $g_0, g_1, \ldots, g_k \in \mathbb{R}^n$ are self-conjugate with respect to the matrix A, then they are linearly independent.

Proof. (By contradiction). Assume that the vectors $g_0, g_1, \ldots, g_k$ are linearly dependent, then without loss of generality, we can assume that the vector g_0 represents as a linear combination of nonzero other vectors, i.e.

$$g_0 = \sum_{i=1}^{k} \alpha_i g_i, \quad \alpha_i \in \mathbb{R} \quad \forall i \in [1, \ldots, k], \quad \sum_{i=0}^{k} \alpha_i^2 \neq 0.$$

Hence we have

$$\langle Ag_0, g_0 \rangle = \left\langle A \sum_{i=1}^{k} \alpha_i g_i, g_0 \right\rangle = \sum_{i=1}^{k} \alpha_i \langle Ag_i, g_0 \rangle = 0.$$

This contradiction proves the lemma.

Thus, if we know the set of mutually conjugate vectors $g_0, g_1, \dots, g_{n-1} \in \mathbb{R}^n$, with respect to a positive definite matrix A, then these vectors can be taken as the basis of a linear space $\mathbb{R}^n$. Consider a quadratic function with a positive definite matrix A

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle, \quad x, b \in \mathbb{R}^n. \tag{1}$$

The function f is strongly convex. Obviously, the point $x^* = -A^{-1}b$ is a minimum of f on $\mathbb{R}^n$.

Let the vectors $g_0, g_1, \ldots, g_{n-1} \in \mathbb{R}^n$ be self-conjugate with respect to the matrix A.

Formulate an algorithm for minimizing a function f on $\mathbb{R}^n$.

Obviously, that the point $x^* = -A^{-1}b$ is the point of minima of f on $\mathbb{R}^n$.

Step 1. Take an initial point $x_0 \in \mathbb{R}^n$. If $x_0 = x^*$, then stop.

Otherwiser calculate

$$\alpha_0 = \arg\min_{\alpha} f(x_0 + \alpha g_0), \quad x_1 = x_0 + \alpha_0 g_0.$$

Step 2. Let the point $x_k,\ k < (n-1)$ be found. If $x_k = x^*$, then stop.

Otherwise calculate

$$\alpha_k = \arg\min_{\alpha} f(x_k + \alpha g_k).$$

Define a new point by the formula

$$x_{k+1} = x_k + \alpha_k g_k.$$

Step 3.

a) If $x_{k+1} = x^*$, then this point is a minimizer and stop.

б) If $k + 1 = n$, then the minimum point is found and the process is stopped.

Otherwise go to step 2.

Theorem 1.

As a result of this algorithm the minimum point of the function f on $\mathbb{R}^n$ is found not more than for n steps.

Proof. Let the point x_n be found. Prove that this point is a minimizer of f on $\mathbb{R}^n$.

On every step

$$\alpha_k = -\frac{\langle f'(x_k), g_k\rangle}{\langle Ag_k, g_k\rangle},$$

and

$$x_n = x_0 + \sum_{k=0}^{n-1} \alpha_k g_k = x_0 - \sum_{k=0}^{n-1} \frac{\langle f'(x_k), g_k\rangle}{\langle Ag_k, g_k\rangle} g_k.$$

Fix an index $i \leq (n-1)$ then

$$x_i = x_0 + \sum_{k=0}^{i-1} \alpha_k g_k.$$

From this we have an equality

$$Ax_i = Ax_0 + \sum_{k=0}^{i-1} \alpha_k A g_k,$$

$$\langle Ax_i, g_j \rangle = \langle Ax_0, g_j \rangle + \sum_{k=0}^{i-1} \alpha_k \langle A g_k, g_j \rangle = \langle Ax_0, g_j \rangle \quad \forall j = 0, \ldots, (i-2).$$

Therefore

$$\langle Ax_i + b, g_j \rangle = \langle Ax_0 + b, g_j \rangle \quad \forall j = 0, \ldots, (i-2).$$

From the last equality we have

$$\langle f'(x_i), g_j \rangle = \langle f'(x_0), g_j \rangle \quad \forall j = 0, \ldots, (i-2),$$

and

$$\langle f'(x_n), g_j \rangle = \langle f'(x_0), g_j \rangle - \frac{\langle f'(x_0), g_j \rangle}{\langle Ag_j, g_j \rangle} \langle Ag_j, g_j \rangle = 0 \quad \forall j = 0, \ldots (n-2).$$

As

$$Ax_n = Ax_0 + \sum_{k=0}^{n-1} \alpha_k A g_k,$$

then

$$\langle Ax_n, g_{n-1}\rangle = \langle Ax_0, g_{n-1}\rangle + \alpha_{n-1}\langle Ag_{n-1}, g_{n-1}\rangle =$$

$$= \langle Ax_0, g_{n-1}\rangle - \frac{\langle f'(x_{n-1}), g_{n-1}\rangle}{\langle Ag_{n-1}, g_{n-1}\rangle}\langle Ag_{n-1}, g_{n-1}\rangle =$$

$$= \langle Ax_0, g_{n-1}\rangle - \langle f'(x_{n-1}), g_{n-1}\rangle =$$

$$= \langle Ax_0, g_{n-1}\rangle - \langle Ax_{n-1} + b, g_{n-1}\rangle = \langle Ax_0, g_{n-1}\rangle - \langle Ax_0, g_{n-1}\rangle - \langle b, g_{n-1}\rangle.$$

Therefore

$$\langle f'(x_n), g_{n-1}\rangle = 0.$$

and

$$\langle f'(x_n), g_j\rangle = 0 \quad \forall j = 0, \ldots (n-1). \tag{2}$$

As the vectors $g_0, g_1, \ldots, g_{n-1}$ form a basis of the space $\mathbb{R}^n$, then from the equality (2) follows that the $f'(x_n) = 0_n$. Consequently, x_n is the minimizer of f on $\mathbb{R}^n$. The theorem is proved.

Corollary. At each step, the point x_k is a minimizer on the linear manifold

$$X = x_0 + \text{lin}\ \{g_0, g_1, \ldots, g_{k-1}\}.$$

Finding a set of conjugate vectors is a separate problem that can be solved in different ways. For example, we can take an arbitrary basis in $\mathbb{R}^n$ and apply the well-known process of the Gram - Schmidt orthogonalization for the scalar product $\langle Ax, x\rangle$.

Conjugate gradient methods

This method belongs to the group of methods of conjugate directions which attempt to locate a local minimum of f. The problem of minimizing a quadratic function of the form (1) is considered. Describe a method for constructing of a system of self-conjugate directions.

As an initial approximation, we choose an arbitrary point $x_0 \in \mathbb{R}^n$.

Step 1. Put

$$g_0 = -f'(x_0), \ \ k = 0.$$

If $f'(x_0) = 0$, then stop.

Step 2. Let the point x_k be found. If $f'(x_k) = 0$, then stop. Define

$$x_{k+1} = x_k + \alpha_k g_k,$$

where

$$\alpha_k = -\frac{\langle f'(x_k), g_k\rangle}{\langle Ag_k, g_k\rangle}.$$

The value of the step-size α_k is chosen from the condition of minimization of quadratic function (1) on the line with a direction vector g_k.

$$g_{k+1} = -f'(x_{k+1}) + \beta_k g_k,$$

$$\beta_k = \frac{\langle f'(x_{k+1}), Ag_k\rangle}{\langle Ag_k, g_k\rangle}. \tag{3}$$

The value of β_k is chosen so that the direction g_k was an A - conjugate with all the previously constructed direction.

Step 3 Put the $k = k + 1$. Go to step 2.

Show that the constructed vectors $g_0, g_1, \dots,$ are self-conjugate with respect to the matrix A.

Lemma 2.

Under the above assumptions the vectors $f'(x_0), f'(x_1), \dots,$ are mutually orthogonal, and the vectors $g_0, g_1, \dots,$ are self-conjugate.

Proof. So

$$f'(x_{k+1}) = f'(x_k) + \alpha_k A g_k,$$

then

$$f'(x_{k+1}) = f'(x_k) - \frac{\langle f'(x_k), g_k\rangle}{\langle A g_k, g_k\rangle} A g_k.$$

From this follows that

$$\langle f'(x_{k+1}), f'(x_k)\rangle = \langle f'(x_k), f'(x_k)\rangle - \frac{\langle f'(x_k), g_k\rangle}{\langle A g_k, g_k\rangle}\langle A g_k, f'(x_k)\rangle.$$

$$\langle f'(x_k), g_k\rangle = -\langle f'(x_k), f'(x_k)\rangle + \beta_{k-1}\langle f'(x_k), g_{k-1}\rangle = -\langle f'(x_k), f'(x_k)\rangle.$$

Besides

$$\langle Ag_k, g_k\rangle = -\langle Ag_k, f'(x_k)\rangle + \beta_{k-1}\langle Ag_k, g_{k-1}\rangle = -\langle Ag_k, f'(x_k)\rangle.$$

From these equations follows that the

$$\langle f'(x_{k+1}), f'(x_k)\rangle = 0.$$

So the gradients at neighboring points of the sequence $\{x_k\}$ are orthogonal. Therefore,

$$f'(x_0)\perp f'(x_1)$$

and by construction the vectors g_0, g_1, are self-conjugate. The induction base is built.

Further assume that the vectors $g_0, \dots, g_k,\ k \le n-1$ are self-conjugate and the vectors $f'(x_0), \dots, f'(x_k)$ are mutually orthogonal.

Then if $j \le k-1$, then

$$\langle f'(x_{k+1}), f'(x_j)\rangle = \langle f'(x_k), f'(x_j)\rangle + \alpha_k \langle A g_k, f'(x_j)\rangle =$$

$$= \alpha_k \langle A g_k, f'(x_j)\rangle = \alpha_k \langle A g_k, -g_j + \beta_{j-1} g_{j-1}\rangle = 0.$$

Orthogonality of gradients $f'(x_k),\ k = 0, .., n-1$, is proved.

As for a quadratic function

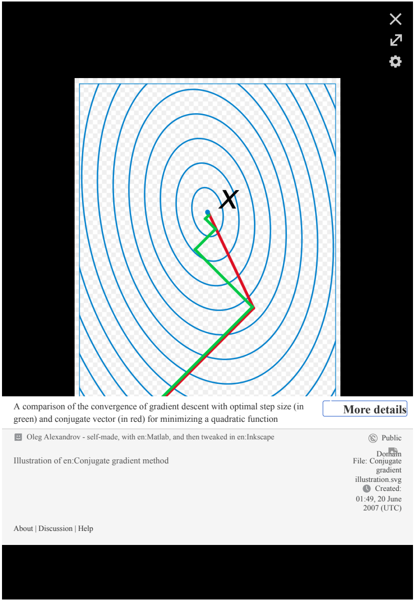
$$f'(x_{j+1}) - f'(x_j) = \alpha_j A g_j,$$

then for $j \le k$, we have

$$\langle A g_j, g_{k+1} \rangle = \langle A g_j, -f'(x_{k+1}) + \beta_k g_k \rangle =$$

$$= -\langle A g_j, f'(x_{k+1}) \rangle = -\left\langle \frac{f'(x_{j+1}) - f'(x_j)}{\alpha_j}, f'(x_{k+1}) \right\rangle = 0$$

The lemma is proved.



A comparison of the convergence of gradient descent with optimal step size (in green) and conjugate vector (in red) for minimizing a quadratic function

Oleg Alexandrov - self-made, with en:Matlab, and then tweaked in en:Inkscape

Public Domain

Illustration of en:Conjugate gradient method

File: Conjugate gradient illustration.svg

Created: 01:49, 20 June 2007 (UTC)

Thus the formulated method is the method of conjugate directions. Therefore, using it we can find the minimizer of quadratic function (1) no more than for n steps.

The conjugate gradient method can be applied for solving the system of linear equations with a positive definite matrix. The method ensures convergence in a finite number of steps and the required accuracy can be achieved more earlier.

The main problem lies in the fact that due to the accumulation of errors orthogonality of basis vectors may be violated that degrades the convergence.

Any convex continuously differentiable function in a neighborhood of a minimizer well approximated by a quadratic, so the methods of conjugate directions successfully for minimizing not only quadratic functions. In this case, the methods cease to be finite and are iterative.

Fletcher and Reeves developed a conjugate gradient method to the case of non-quadratic functions. To use it for smooth functions, it is necessary to transform formula (3) so that it does not use a matrix A.

There are many modifications of the conjugate gradient method for arbitrary smooth functions.

As a rule, methods differ in the choice of the coefficient β_k.

Formulate *the Fletcher- Reeves algorithm.*

Step 1. Take an initial point $x_0 \in \mathbb{R}^n$.

If $f'(x_0) = 0_n$, then x_0 is a minimizer and stop.

If $f'(x_0) \neq 0$, then calculate

$$\alpha_0 = \arg\min_{\alpha} f(x_0 + \alpha g_0), \quad x_1 = x_0 + \alpha_0 g_0.$$

Put $g_0 = -f'(x_0)$.

Step 2. Let the point x_k be found. If $f'(x_k) = 0$, then x_k is a minimizer and stop.

Otherwise calculate

$$\alpha_k = \arg\min_{\alpha} f(x_k + \alpha g_k),$$

$$x_{k+1} = x_k + \alpha_k g_k,$$

$$g_{k+1} = -f'(x_{k+1}) + \beta_k g_k,$$

$$\beta_k = \frac{||f'(x_{k+1})||^2}{||f'(x_k)||^2}. \tag{4}$$

Step 3. Checking the stopping criterion.

If the criterion is not satisfied then go to Step 2.

You can also choose the coefficient β_k by the formula

$$\beta_k = \frac{\langle f'(x_{k+1}), f'(x_{k+1}) - f'(x_k)\rangle}{||f'(x_k)||^2}. \tag{5}$$

Method with such choice of the coefficient β_k is called the Polak - Reiber method.

If the function f is a quadratic with a positive definite matrix A, then the methods of Fletcher - Reeves and Polak - Ribier give the same result.

Methods of conjugate directions sensitive to errors arising in the calculation process.

In the practical realization of methods of conjugate directions computational errors can be lead to what the vectors g_k will cease to indicate the direction of descent and the convergence of the method can be broken.

Therefore under using such methods the procedure of "recovery" is carried out.

After a certain number of steps the method is updating, i.e. the found point x_k is considered as a new initial approximation.

Theorem 2.

Let a continuous differentiable function f is bounded below on $\mathbb{R}^n$, $x_0 \in \mathbb{R}^n$ is an initial point and gradients f' satisfy a Lipschitz condition

$$||f'(x) - f'(y)|| \leq L||x - y|| \quad \forall x, y \in \mathbb{R}^n, \; L > 0.$$

Then

$$||f'(x_k)|| \xrightarrow[k \to \infty]{} 0.$$

For a strongly convex smooth functions satisfying certain additional constraints a sequence of points $\{x_k\}$ generated by the algorithm converges to a minimizer x^* with superlinear rate.

At each iteration the Fletcher-Reeves or Polak-Reiber methods once is calculated value of the function and its gradient and is solved a problem of one-dimensional optimization.

Thus the complexity of one step of the conjugate gradient method is of the same order as the complexity of a step of the method of steepest descent.

In practice the conjugate gradient method shows the better convergence rate.

Methods in which a new approach depends on s the previous approaches are called s -step methods.

In gradient methods at each step x_k the information obtained in the previous iterations is not used, i.e, they are one-step methods. Methods of conjugate directions are two-step ones, as on k-th step the information about the points x_k and x_{k+1} is used in the construction of the direction of descent g_k.