Unconstrained optimization
The Newton method
The quasi-Newton methods

31 октября 2017 г.

Quasi-Newton methods for optimization combine the advantages of the method of steepest descent and Newton's method.
They are based on the idea of approximating of the inverse Hessian matrix of the function which it is necessary to be minimized.

These methods are relaxation.

They have a common structure in the construction of an iterative sequence $\{x_k\}$. A point x_{k+1} is calculated on $k+1$ step by the rule:

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

where a positive definite matrix H_k is recalculated by the recurrence method so to satisfy the condition

$$H_k - f''(x_k) \xrightarrow[k\to\infty]{} 0. \tag{1}$$

Let us denote

$$p_k = -H_k f'(x_k), \quad y_k = f'(x_{k+1}) - f'(x_k).$$

Decompose the gradient of the minimized function f' in the Taylor series in the neighborhood of the point x_k. We have

$$f'(x_k) - f'(x_{k+1}) = f''(x_{k+1})(x_k - x_{k+1}) + o(||x_k - x_{k+1}||).$$

$$\left[f''(x_{k+1})\right]^{-1}(f'(x_k) - f'(x_{k+1})) = x_k - x_{k+1} + o(||x_k - x_{k+1}||).$$

In order to condition (1) is satisfied it is necessary that

$$H_{k+1}y_k = \alpha_k p_k.$$

This condition is called *the quasi-Newton condition*. It is decisive in the construction of these algorithms. Consider the problem of calculating matrices H_{k+1}.
For their calculation the following lemma from linear algebra is used.

Lemma 1 . *Let A be a non-singular matrix of order $(n \times n)$, $B = ab^T$ is a matrix rank 1 and $\langle A^{-1}a, b\rangle \neq -1$. Then the equality*

$$(A + B)^{-1} = A^{-1} - \frac{A^{-1}a(A^{-1}b)^T.}{(1 + \langle A^{-1}a, b\rangle)}.$$

holds.

Since the Hessian matrix is symmetric, it is necessary to maintain symmetry. Specify some conversion formulas

1. The Davidon-Fletcher-Powell formula (or DFP; named after William C. Davidon, Roger Fletcher, and Michael J. D. Powell):

$$H_{k+1} = H_k - \frac{H_k y_k (y_k)^T H_k}{\langle H_k y_k, y_k \rangle} + \alpha_k \frac{p_k (p_k)^T}{\langle p_k, y_k \rangle}, \quad H_0 > 0.$$

2. The Broyden formula:

$$H_{k+1} = H_k - \frac{(\alpha_k p_k - H_k y_k)(\alpha_k p_k - H_k y_k)^T}{\langle \alpha_k p_k - H_k y_k, y_k \rangle}, \quad H_0 > 0.$$

3. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula:

$$H_{k+1} = H_k + \frac{\beta_k p_k (p_k)^T - p_k (y_k)^T H_k - H_k y_k (p_k)^T}{\langle y_k, p_k \rangle},$$

$$\beta_k = \alpha_k + \frac{\langle H_k y_k, y_k \rangle}{\langle y_k, p_k \rangle}, \quad H_0 > 0.$$

As you can be seen from these formulas, at each iteration a matrix of rank 1 is added to the matrix H_k.

As a rule H_0 is chosen as the identity matrix if no special considerations.

Step size α_k is often defined as one that minimizes the objective function along a selected direction, that is, it is necessary to solve the problem of one dimensional minimization at each iteration

$$\alpha_k = \arg\min_{\alpha > 0} f(x_k + \alpha p_k).$$

Sometimes the parameter α_k is chosen at the each iteration by bisection the step-size.

Formulate the "fundamental algorithm"of quasi-Newton methods.

Step 1. Take an initial point $x_0 \in \mathbb{R}^n$.
If $f'(x_0) = 0_n$, then x_0 is a minimizer of f and stop.
If $f'(x_0) \neq 0$, then put

$$H_0 = H, \;\; p_0 = -H_0 f'(x_0),$$

where H_0 is a positive definite matrix, and calculate

$$\alpha_0 = \arg\min_{\alpha} f(x_0 + \alpha p_0), \quad x_1 = x_0 + \alpha_0 p_0.$$

Step 2 Let x_k is constructed. If $f'(x_k) = 0$, then x_k is a minimizer of f and stop. Otherwise calculate

$$p_k = -H_k f'(x_k),$$

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

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

Step 3. Check the stopping criterion. If the criterion is not satisfied, then $k = k + 1$ and go to step 2.

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for quadratic functions is finite (goes to the exact solution is not more than n steps).

For non-quadratic functions under fairly general conditions it has a locally superlinearly convergent.

The number of different variations and modifications quasi-Newton methods is very large.

Newton's method

Second-order methods are developed to improve the convergence of gradient methods and other methods the first-order. One of such method is the method of Newton.

Originally Newton's method was intended for solving equations. However, it can be also applied to solutions of problems of an unconstrained and a constrained optimization.

Newton's method of minimization function is a generalization of the well-known method of Newton for finding the root of the equation

$$f(x) = 0.$$

This method is a method of the second order since for its implementation the Hessian matrix which is the matrix of second derivatives is used.

Methods of the second order using the information about of the second derivative are based on a quadratic approximation of the objective function.

The idea is to replace the function f in the neighborhood of the current point x_k its quadratic approximation.

Let a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ be given.

Expand the function f in the Taylor series in the neighborhood of x_k

$$f(x) = f(x_k) + \langle f'(x_k), x - x_k\rangle +$$

$$+\frac{1}{2}\langle f''(x_k)(x - x_k), x - x_k\rangle + o(||x - x_k||^2) \quad x, x_k \in \mathbb{R}^n.$$

Put away all the expansion terms above the second order.

Denote

$$\varphi(x) = f(x_k) + \langle f^{'}(x_k), x - x_k\rangle + \frac{1}{2}\langle f''(x_k)(x - x_k), x - x_k\rangle.$$

The function φ approximates f at x. If the Hessian matrix f'' is positive definite (hence it is nonsingular) then the function φ attains its minimum on $\mathbb{R}^n$ in an unique point.

Compute its gradient and equate it to zero.

$$\varphi'(x) = f'(x_k) + f''(x_k)(x - x_k) = 0_n.$$

From here

$$x - x_k = -\left[f''(x_k)\right]^{-1} f'(x_k).$$

So for the new estimate of the solution we can take the point

$$x_{k+1} = x_k - \left[f''(x)\right]^{-1} f'(x).$$

It is easy to see that this formula coincides with the classical Newton method for solving systems of nonlinear equations

$$f'(x) = 0_n.$$

Note that here both the direction of movement and the step size are fixed.

In the problem of finding the minimum of a quadratic function with a positive definite matrix A Newton's method converges in one step, regardless of the initial point. In fact, if

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c, \quad x, b \in \mathbb{R}^n, \quad A > 0,$$

then $x^* = -A^{-1}b$ is the minimizer of f.

Let us take any initial point $x_0 \in \mathbb{R}^n$. The vector

$$y = -A^{-1}(Ax_0 + b) = -x_0 - A^{-1}b.$$

is a descent direction.

Then

$$x^* = x_0 + (-x_0 - A^{-1}b) = -A^{-1}b.$$

This is because the direction of descent

$$g(x_0) = -A^{-1}f'(x_0)$$

at any point x_0 coincides with the direction of a minimum point x^*. If the function f is not quadratic but convex, then Newton's method guarantees its monotonic decrease from point to point.

Formulate an "fundamental algorithm"of the Newton's method with a constant step which is equal to 1.

Step 1. A starting point $x_0 \in \mathbb{R}^n$ is chosen from some neighborhood of an minimizer. If $f'(x_0) = 0_n$, then the point x_0 is a stationary point of f. The process is completed.

Step 2. Let a point $x_k \in \mathbb{R}^n$ have been found. Define

$$x_{k+1} = x_k - \left[f''(x_k)\right]^{-1} f'(x_k). \tag{2}$$

If $f'(x_{k+1}) = 0_n$, then the point x_{k+1} is a stationary point of f. The process is completed.

Step 3 Put $k = k + 1$ and go to step 2.

This iterative process is constructing a sequence of points $\{x_k\}$, which under certain conditions converges to a stationary point x^* of f with a quadratic rate i.e., to the point where

$$f'(x^*) = 0_n.$$

Theorem 1.

Let a function $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, strongly convex with strong convexity constant m, the Hessian matrix satisfies the Lipschitz condition, i.e., there exists a constant $L > 0$ such that the inequality

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

holds, an initial point x_0 be taken from the condition

$$L||f'(x_0)|| \leq 8m^2 q, \quad q \in (0, 1).$$

Then the sequence of points $\{x_k\}$ constructed by formulas (2), converges with quadratic rate to the minimizer x^* of f and

$$||x_k - x^*|| \leq \left(\frac{4m^2}{L}\right) q^{2^k}.$$

Proof. As the function f is strongly convex then its the Hessian matrix is positive definite

$$\langle f''(x)g, g\rangle \geq 2m||g||^2 \quad g \in \mathbb{R}^n. \tag{3}$$

Therefore f'' is inverse. From the properties of convex functions it is known that

$$\langle f'(x) - f'(y), x - y\rangle \geq 2m||x - y||^2 \quad \forall x, y \in \mathbb{R}^n. \tag{4}$$

If x^* is the minimizer of f then $f'(x^*) = 0_n$.

Therefore from (4) and from the Cauchy-Schwarz inequality we have

$$2m||x^k - x^*||^2 \leq ||f'(x_k)||||x_k - x^*||.$$

Thence

$$||x^k - x^*|| \leq \frac{1}{2m}||f'(x_k)||. \tag{5}$$

As

$$f'(x_{k+1}) = f'(x_k) + \int_0^1 f''(x_k + t(x_{k+1} - x_k))\,(x_{k+1} - x_k)\,dt,$$

then

$$f'(x_{k+1}) - f'(x_k) - f''(x_k)\,(x_{k+1} - x_k) =$$

$$= \int_0^1 f''(x_k + t(x_{k+1} - x_k))\,(x_{k+1} - x_k)\,dt - f''(x_k)\,(x_{k+1} - x_k) =$$

$$= \int_0^1 \left(f''(x_k + t(x_{k+1} - x_k)) - f''(x_k)\right)\,(x_{k+1} - x_k)\,dt.$$

Using of the Lipschitz property of the Hessian matrix we have

$$||\int_0^1 \left(f''(x_k + t(x_{k+1} - x_k)) - f''(x_k)\right) (x_{k+1} - x_k)\, dt|| \leq$$

$$\leq L \int_0^1 t||x_{k+1} - x_k||^2\, dt = \frac{L}{2}||x_{k+1} - x_k||^2.$$

From here

$$||f'(x_{k+1}) - f'(x_k) - f''(x_k)\,(x_{k+1} - x_k)|| \leq \frac{L}{2}||x_{k+1} - x_k||^2.$$

As

$$f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0_n,$$

then

$$||f'(x_{k+1})|| \leq \frac{L}{2}||x_{k+1} - x_k||^2 \leq \frac{L}{2}|| \left[f''(x_k) \right]^{-1} ||^2 ||f'(x_k)||^2 \leq$$

$$\leq \frac{L}{8m^2}||f'(x_k)||^2,$$

since

$$|| \left[f''(x_k) \right]^{-1} || \leq \frac{1}{2m}.$$

Easy to observe that

$$||f'(x_k)|| \leq \frac{L}{8m^2}||f'(x_{k-1})||^2 \leq \left(\frac{L}{8m}\right)^3 ||f'(x_{k-2})||^4 \leq \dots$$

$$\dots \leq \left(\frac{L}{8m}\right)^{2^k-1} ||f'(x_0)||^{2^k} == \left(\frac{8m^2}{L}\right) q^{2^k}.$$

Thence and from (5) it follows

$$||x_k - x^*|| \leq \left(\frac{4m^2}{L}\right) q^{2^k}.$$

The theorem is proved.

Essential disadvantages of Newton's method are:

1) the complexity of the choice of the initial approximation x_0 in a small neighborhood of the desired minimum x^*;

2) at each iteration it requires the computation of the gradient and the inverse of Hessian matrix. This can be quite cumbersome.

Note that the Newton algorithm is formulated above, nowhere uses the objective function.

You can slightly modify the method.
Select a starting point x_0, compute the Hessian matrix f'' and find the reverse matrix thereto. If the point x_k is already defined, then the next point x_{k+1} is computed by the formula

$$x_{k+1} = x_k - [f''(x_0)]^{-1} f'(x_k).$$

Thus, the computational complexity of the method is reduced.

It will converge under the same conditions as previously stated, however, the fee for simplification is high.

The method with the selection step converges linearly, rather than a quadratic rate.

There may be some compromises.

For example, to calculate the matrix $[f'']^{-1}$ not at each iteration, and in a few.

There are several ways to modify Newton's method, to achieve global convergence.

As a rule, they are designed to save ґ the main advantage of the Newton method:

its good convergence, reduce labor intensity and reduce the demand the choice of the initial approximation.

Consider one of the ways.

First, since the approximation of f by the function φ occurs only in the neighborhood of x_k, then we can affect on a step-size.

Newton's method with a step regulation

Note that Newton's algorithms formulated above never uses an objective function.
Consider the algorithm in which step-size will be chose according to some rules.
These methods are called *the Newton-Raphson methods.* They are constructed by analogy with the gradient methods with a variable step.

Choose an arbitrary starting point of $x_0 \in \mathbb{R}^n$.

Check the stopping condition.

Suppose that we have found a point $x_k \in \mathbb{R}^n$.

If the stop condition is not satisfied, then the iterative sequence $\{x_k\}$ is constructed according to the rule

$$x_{k+1} = x_k - \alpha_k \left[f''(x_k)\right]^{-1} f'(x_k), \quad \alpha_k > 0. \tag{6}$$

Choosing a step-size α_k in (6) is produced from a one-dimensional minimization conditions or the Armijo rules, where the initial step α is taken from the interval $(0, 1]$, and then crushed until at some α_k the next inequality

$$f(x_k - \alpha_k \left[f''(x_k)\right]^{-1} f'(x_k)) - f(x_k) \leq$$

$$\leq -\varepsilon\alpha_k \langle f'(x_k), \left[f''(x_k)\right]^{-1} f'(x_k)\rangle, \quad \varepsilon \in (0, 1).$$

is fulfilled.

We can prove that the Newton - Raphson method for strongly convex functions has a quadratic rate of convergence to a single minimum point if a starting point is in the neighborhood of it and if a starting point is away from the minimizer it converges linearly.

Remark.

In the classical Newton's method $\alpha_k = 1$.