Unconstrained optimization
Gradient methods

25 сентября 2017 г.

Descend methods for smooth optimization

Many methods of descent fit into the following scheme

The initial step.

Take an initial point $x_0 \in \mathbb{R}^b$ and a given accuracy of calculation $\varepsilon > 0$.

Let a point $x_k \in \mathbb{R}^n$ have been computed.

Step 1. (Checking of stopping criteria).

Check of stopping criteria. If they are satisfied, the calculation is stopped and x_k is taken as a solution .

Otherwise, go to the next step.

Step 2. Select of a descent direction (find the vector $g_k \in \mathbb{R}^n$).

Select a step-size. It is necessary to define $\alpha_k > 0$ for which

$$f(x_k + \alpha_k g_k) < f(x_k).$$

Step 4. Choose a new estimate of the solution. Put

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

and go to step 1.

Descent methods are *relaxation* because at each step the value of the function decreases, i.e.

$$f(x_{k+1}) < f(x_k).$$

The rate of convergence is the main characteristic of numerical optimization methods. The higher the rate of convergence, the less iterations have to be done to achieve the given accuracy.

It is said that the sequence $\{x_k\}$ converges linearly to x^* (linear convergence) if

$$||x_{k+1} - x^*|| \leq q||x_k - x^*||, \quad 0 < q < 1. \tag{1}$$

It is said that the sequence $\{x_k\}$ converges superlinearly to x^* if

$$||x_{k+1} - x^*|| \leq q^k||x_k - x^*||, \quad 0 < q_k < 1, \quad q_k \to 0. \tag{2}$$

It is said that the sequence $\{x_k\}$ has the quadratic convergence to x^* if $\exists C > 0$ and

$$||x_{k+1} - x^*|| \leq C||x_k - x^*||^2, \quad \text{при } k \geq K, \ C > 0. \tag{3}$$

Sometimes, keeping the same terminology inequalities (1) - (3) are replaced by the inequality:

$$||x_{k+1} - x^*|| \leq C_1 q_{k+1}, \quad \text{for } k \geq K, \ C_1 > 0,$$

$$||x_{k+1} - x^*|| \leq C_2 q_{k+1} q_k \dots q_1, \quad \text{for } k \geq K, \ C_2 > 0.$$

$$||x_{k+1} - x^*|| \leq C_3 q^{2^{k+1}}, \quad \text{при } k \geq K, \ C_3 > 0.$$

If a direction of descent $g_k \in \mathbb{R}^n$ for a function f at $x_k \in \mathbb{R}^n$ is defined then a step size α_k can be defined by several methods.

1. One-dimensional optimization:

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

2. One-dimensional optimization with constraints:

$$\alpha_k = \arg\min_{\alpha\in(0;\bar{\alpha}]} f(x_k + \alpha g_k), \quad \bar{\alpha} > 0.$$

3. Bisection algorithm (Armijo rule): for fixed values of the numerical parameters $\alpha \in (0;1)$, $\beta \in (0;0.5)$, find the first value of $i = 0, 1, \ldots,$ for which the inequality

$$f(x_k + \alpha^i g_k) \le f(x_k) + \beta\alpha^i \langle f'(x_k), g_k)\rangle$$

holds. Usually $\alpha = \dfrac{1}{2}$.

In practice, we need to run the bisection algorithm with a stopping criterion. Some relevant stopping criteria are:

- stop after a fixed number of iterations;
- stop when the interval becomes small;
- stop when $||f'(x_k)||$ becomes small.

4. Constant step $\alpha_k = \alpha$.

Example 1. Consider a quadratic function with a positive definite matrix $A > 0$

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

Fix points $x, g \in \mathbb{R}^n$ and define a function

$$\psi(\alpha) = f(x + \alpha g), \ \alpha \in \mathbb{R}.$$

As the function f is strongly convex, then the function ψ is also strongly convex. Calculate

$$\min_{\alpha \in \mathbb{R}} \psi(\alpha). \tag{4}$$

A minimizer α^* of this function is unique. Find it.

Calculate $\psi'(\alpha)$.

$$\psi'(\alpha) = \langle f'(x + \alpha g), g\rangle = \langle A(x + \alpha g) + b, g\rangle =$$

$$= \langle Ax + \alpha Ag + b, g\rangle = \langle f'(x) + \alpha Ag, g\rangle = \langle f'(x), g\rangle + \alpha\langle Ag, g\rangle.$$

As $\psi'(\alpha^*) = 0$, then, if $f'(x) \neq 0_n$, we have

$$\alpha^* = -\frac{\langle f'(x), g\rangle}{\langle Ag, g\rangle}. \tag{5}$$

Under considering gradient methods we assume that gradients and Hessian matrices can be calculated with sufficient accuracy.

As a rule an analytical expression for the gradient is very difficult to obtain in many problems encountered in practice.

It is advisable to develop a scheme for the approximate calculation of gradients.

Approximation of a gradient by one-sided arguments is the simplest variant of this scheme.

Such approximation is directly based on the definition partial derivatives and for sufficiently small values of the increments of variables gives a very accurate assessment.
The choice of this increment is dependent on a form of the function f.
Due to the additional calculation of the function we can be improved the accuracy of the approximation by using of bilateral increments of arguments (central finite difference).

Here are some formulas for numerical differentiation:

1) $y' \approx \dfrac{f(x+h)-f(x)}{h}$,

2) $y' \approx \dfrac{f(x+h)-f(x-h)}{2h}$,

3) $y' \approx \dfrac{-f(x+2h)+8f(x+h)-8f(x-h)+f(x-2h)}{12h}$,

4) $y' \approx \dfrac{-f(x+2h)+4f(x+h)-3f(x)}{2h}$,

5) $y' \approx \dfrac{f(x+3h)-6f(x+2h)+18f(x+h)-10f(x)-3f(x-h)}{12h}$,

6) $y'' \approx \dfrac{f(x+h)-2f(x)+f(x-h)}{h^2}$,

7) $y'' \approx \dfrac{-f(x+3h)+4f(x+2h)-5f(x+h)+2f(x)}{h^2}$.

Gradient methods

Gradient methods are used to optimize continuously differentiable functions.

They belong to a group of the first order one.

Among the optimization methods they are the easiest for realization.

Gradient methods for unconstrained optimization practically consist of determining the stationary point of an objective function, i.e. it is necessary to find $x^* \in \mathbb{R}^n$ at which $f'(x^*) = 0$.

Consider the expansion of a continuously differentiable function at a point x

$$f(x + \alpha g) = f(x) + \alpha \langle f'(x), g \rangle + o(\alpha).$$

By virtue of this expansion, it is easy to see, that any direction g, $g \neq 0_n$, for which

$$\langle f'(x), g \rangle < 0,$$

is a descent direction of f at x.

Since the gradient shows the direction in which the function grows faster than the whole, it is usually as the descent direction is taken the antigradient.

As for the rate of convergence then under certain conditions for the objective function we can show mainly only the linear rate of the convergence.

Consider some auxiliary lemmas needed for proving the convergence in the future .

Lemma 1.

Let the function f be continuously differentiable on $\mathbb{R}^n$ and its gradients satisfy the Lipschitz condition

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

$y = x - \alpha f'(x), \quad \alpha > 0$. *Then the inequality*

$$f(y) \leq f(x) - \alpha \left(1 - \frac{\alpha L}{2}\right) ||f'(x)||^2. \tag{6}$$

holds.

P r o o f. Denote by $z = -\alpha f'(x)$. Since

$$f(x+z) = f(x) + \int_0^1 \left\langle f'(x+tz), z \right\rangle \, dt =$$

$$= f(x) + \int_0^1 \left\langle f'(x+tz) - f'(x), z \right\rangle \, dt + \langle f'(x), z \rangle.$$

and

$$\left\langle f'(x+tz) - f'(x), z \right\rangle \leq ||(f'(x+tz) - f'(x)|| \cdot ||z|| \leq Lt||z||^2,$$

then

$$f(x+z) \leq f(x) + L||z||^2 \int_0^1 t\,dt + \langle f'(x), z\rangle =$$

$$= f(x) + \frac{\alpha^2 L}{2}||f'(x)||^2 - \alpha||f'(x)||^2 =$$

$$= f(x) - \alpha\left(1 - \frac{\alpha L}{2}\right)||f'(x)||^2.$$

The lemma is proved.

Let the function f be continuously differentiable on $\mathbb{R}^n$, $x_k \in \mathbb{R}^n$, $r > 0$. Then

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

where

$$\frac{o(||x - x_k||)}{||x - x_k||} \xrightarrow[||x - x_k|| \to 0]{} 0.$$

Consider the linear part of this expansion. Denote by

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

The function φ approximates f in the neighborhood of x_k with the accuracy $o(||x - x_k||)$. Find

$$\min_{x \in S(x_k, r)} \langle f'(x_k), x - x_k \rangle,$$

where

$$S(x_k, r) = \{x \in \mathbb{R}^n \mid \ ||x - x_k|| \leq r\}.$$

The linear function $\langle f'(x_k), x - x_k \rangle$ on the sphere $S(x_k, r)$ reaches its minimum value at the point

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

if $f'(x_k) \neq 0_n$. Therefore the point x_{k+1} can be taken for a new estimate of the solution.

If an antigradient is taken as a descent direction of descent and a step-size is defined from the one-dimensional minimization, then such a gradient method is called a method of steepest descent, or the Cauchy method .

Formulate a "basic algorithm" of the steepest descent method.

Step 0. Take an initial point $x_0 \in \mathbb{R}^n$ and $\varepsilon > 0$.

If $f'(x_0) = 0$, then x_0 is a stationary point.

The algorithm stops.

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

Step 1. Calculate the gradient $f'(x_k)$.

If $f'(x_k) = 0$, then x_k is a stationary point.

The algorithm stops.

Otherwise, go to step 2.

Step 2. Determine the step-size from the condition

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

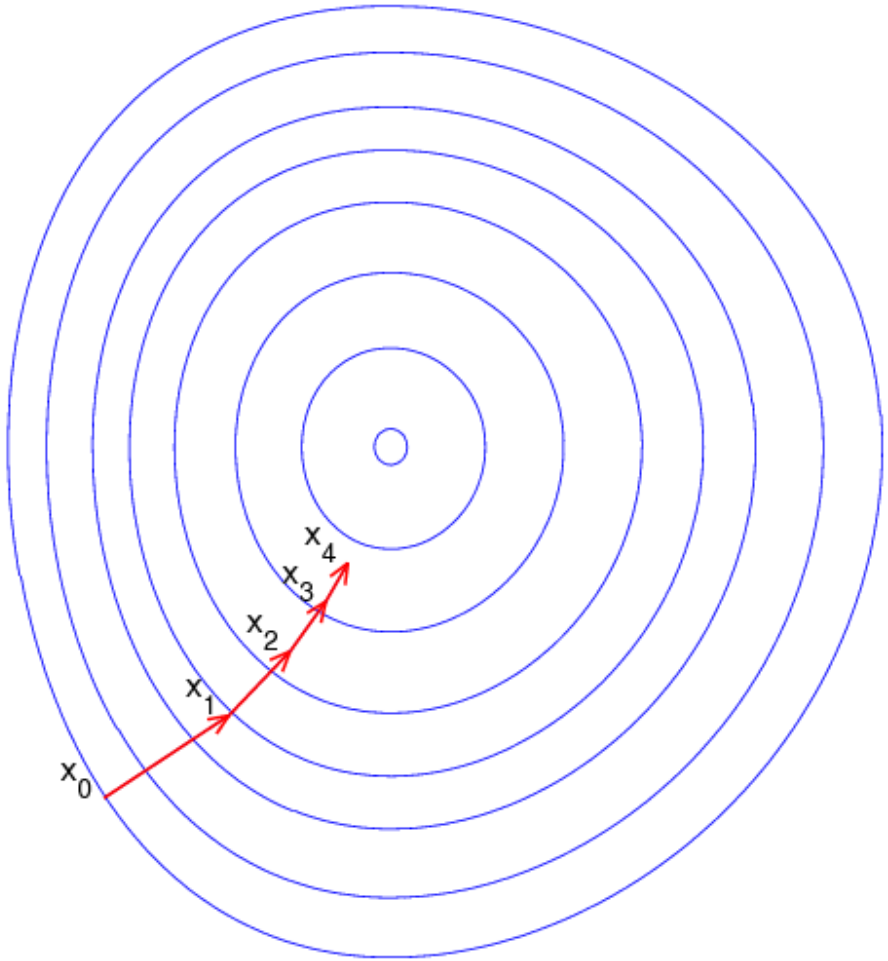
Step 3. Calculate a new iteration point according to the rule:

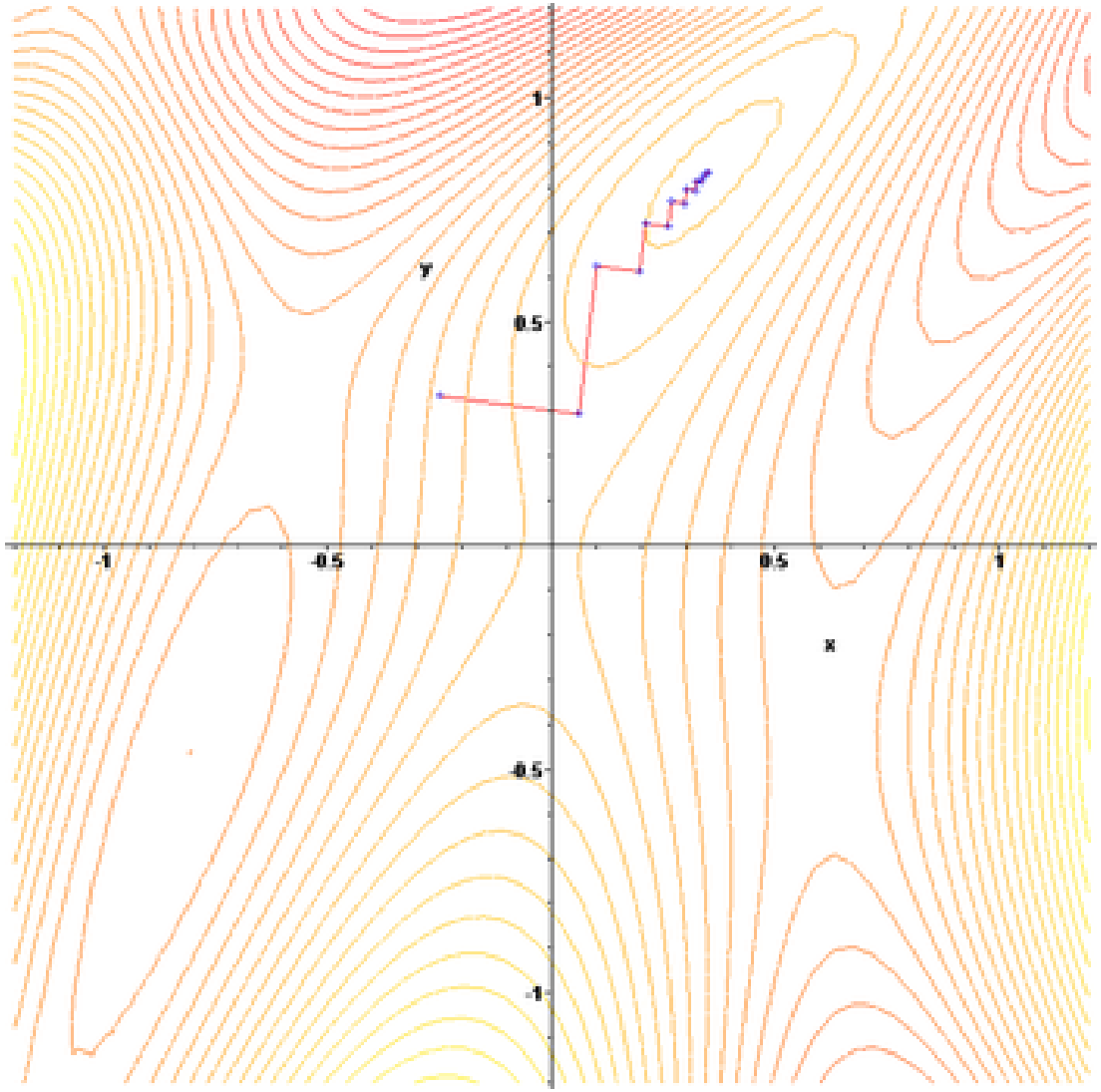
$$x_{k+1} = x_k - \alpha_k f'(x_k).$$

Step 4. Check the stopping criteria. If

$$\max\{||x_{k+1} - x_k||, |f(x_{k+1}) - f(x_k)|, ||f'(x_{k+1})||\} \leq \varepsilon,$$

then the algorithm stops. Otherwise, go back to step 1, putting $k = k + 1$.


x_0
x_1
x_2
x_3
x_4


1
0.5
-0.5
-1
-1
-0.5
0.5
1
x
y

The steepest descent method is stable, i.e. for sufficiently small α_k the inequality

$$f(x_{k+1}) < f(x_k)$$

is satisfied.

Consider some theorems on the convergence of the gradient descent method with the constant step.

Theorem 1.

Let the function f be continuously differentiable, bounded below on $\mathbb{R}^n$ and Lipschitz condition for its gradient is satisfied

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

and

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

where $0 < \alpha < \frac{2}{L}$. *Then for any choice of the initial point* $x_0 \in \mathbb{R}^n$ *the statements*

$$f(x_{k+1}) < f(x_k), \quad \lim_{k \to 0} ||f'(x_k)|| = 0.$$

are true.

P r o o f. As

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

then from (6) it follows that

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

Denote by

$$\theta = \alpha \left(1 - \frac{\alpha L}{2}\right),$$

then

$$f(x_{k+1}) \leq f(x_k) - \theta ||f'(x_k)||^2. \tag{7}$$

By the choice of the step-size $\theta > 0$.

Suppose that $||f'(x_k)||$ doesn't converge to 0. Then, without loss of generality, we can assume, that there exists such positive number a, which satisfies

$$||f'(x_k)|| \geq a > 0 \quad \forall\ k.$$

It follows from this inequality and (7) that

$$f(x_k) \xrightarrow[k\to\infty]{} -\infty.$$

This statement contradicts of the boundedness from below of the objective function.

Corrolary 1. This theorem is also true for steepest descent method.

Corrolary 2. It should be noted, that the sequence $\{x_k\}$ generated by the algorithm can be non convergent, but its any limit point is stationary.

For example, for a function

$$f(x) = \frac{1}{1+x^2}, \ x \in \mathbb{R},$$

the sequence $\{x_k\}$ of ther gradient method with constant step starting with the initial point x_0 tends to $+\infty$

We also note that the described gradient method does not distinguish between local and global points of minima. Therefore, in order to make a conclusion about the convergence of the sequence $\{x_k\}$ to the minimum point x^* it is necessary to impose additional conditions ensuring, in particular, the existence and uniqueness of the solution of our optimization problem. One variant of such conditions will be described below.

Theorem 2.

Let the function f be continuously differentiable, strongly convex with parameter $m > 0$, have a Lipschitz continuous gradient with modulus L,

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

and

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

where $0 < \alpha < \frac{2}{L}$. Then for any initial point $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ tends to the unique minimum point x^* of f with the rate of the geometric progression

$$\lim_{k \to \infty} x_k = x^*, \quad ||x_k - x^*|| \leq q^k ||x_0 - x^*||, \quad q \in [0, 1), \quad C > 0.$$

Proof. As

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

and

$$||f'(x_k)||^2 \geq 4m(f(x_k) - f(x^*)),$$

then

$$f(x_{k+1}) \leq f(x_k) - 2m\alpha(2 - L\alpha)(f(x_k) - f(x^*),$$

$$f(x_{k+1}) - f(x^*) \leq (1 - 2m\alpha(2 - L\alpha))(f(x_k) - f(x^*)).$$

Put $q_0 = 1 - 2m(2 - L\alpha)$. Then

$$f(x_{k+1}) - f(x^*) \leq q_0 \left(f(x_k) - f(x^*)\right).$$

If $k = 0$ then

$$f(x_1) - f(x^*) \leq q_0 \left(f(x_0) - f(x^*)\right).$$

Since

$$f(x_0) > f(x^*), \quad f(x_1) > f(x^*),$$

then $0 < q_0 < 1$.

If $k = 1$ then

$$f(x_2) - f(x^*) \leq q_0(f(x_1) - f(x^*)) \leq q_0^2(f(x_0) - f(x^*)).$$

Thus

$$f(x_{k+1}) - f(x^*) \leq q_0^{k+1}(f(x_k) - f(x^*)). \tag{8}$$

Since

$$q_0^{k+1} \xrightarrow[\alpha \to 0]{} 0,$$

then

$$f(x_k) \to f(x^*).$$

From the properties of strongly convex functions, we have

$$f(x_k) \geq f(x^*) + \langle f'(x^*), x_k - x^* \rangle + m||x_k - x^*||^2.$$

As $f'(x^*) = 0$, then

$$f(x_k) \geq f(x^*) + m||x_k - x^*||^2.$$

From this inequality and from (8) we have

$$||x_k - x^*||^2 \leq \frac{1}{m} q^k (f(x_0) - f(x^*)).$$

Let $C = \sqrt{\frac{f(x_0) - f(x^*)}{m}}$, $q = \sqrt{q_0}$.

Then

$$||x_k - x^*|| \leq C\, q^k.$$

Thus the method has a linear rate of convergence. From this inequality we also obtain

$$x_k \xrightarrow[k \to +\infty]{} x^*.$$

The theorem is proved.

Theorem 3.

Let the function f be twice continuously differentiable, strongly convex on $\mathbb{R}^n$ with parameter $m > 0$, its matrix of second derivatives satisfies

$$\left\langle f''(x)g, g, \right\rangle \leq M||g||^2 \quad \forall g \in \mathbb{R}^n, \ \forall g \in \mathbb{R}^n,$$

and a step-size chosen from the condition of one-dimensional minimization. Then for an arbitrary initial point $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ converges to a minimum point x^* of f at a rate of geometric progression

$$f(x_k) - f(x^*) \leq q^k(f(x_0) - f(x^*)), \quad ||x_k - x^*|| < C(\sqrt{q})^k,$$

where $q \in (0, 1)$, $C > 0$.

These methods are good for their reliability: at each step the function is decreased, and in the limit under fairly broad conditions the achievement of minimum is guaranteed.
Usually, it is advisable to apply the gradient descent methods during the initial stages of minimization using the found result under a relatively small number of iterations the value x_k as the initial approximation for more complex methods which have a higher rate of convergence, for example, Newton's method.

Gradient methods converge poorly for functions for which the matrix of second derivatives (Hessian) is ill-conditioned. In this case maximum and minimum eigenvalues are strongly differing and level-curve of function to be minimized is greatly elongated.
An antigradient direction deviates significantly from directions to the minimum point, which leads to slow the rate of convergence
The rate of convergence of gradient methods also depends on the accuracy of calculation of the gradient and the step-size of descent. Loss of accuracy, and it usually occurs in the vicinity of minimum points or gully situation may even disrupt the convergence of the gradient descent.
Method starts to be cyclical.

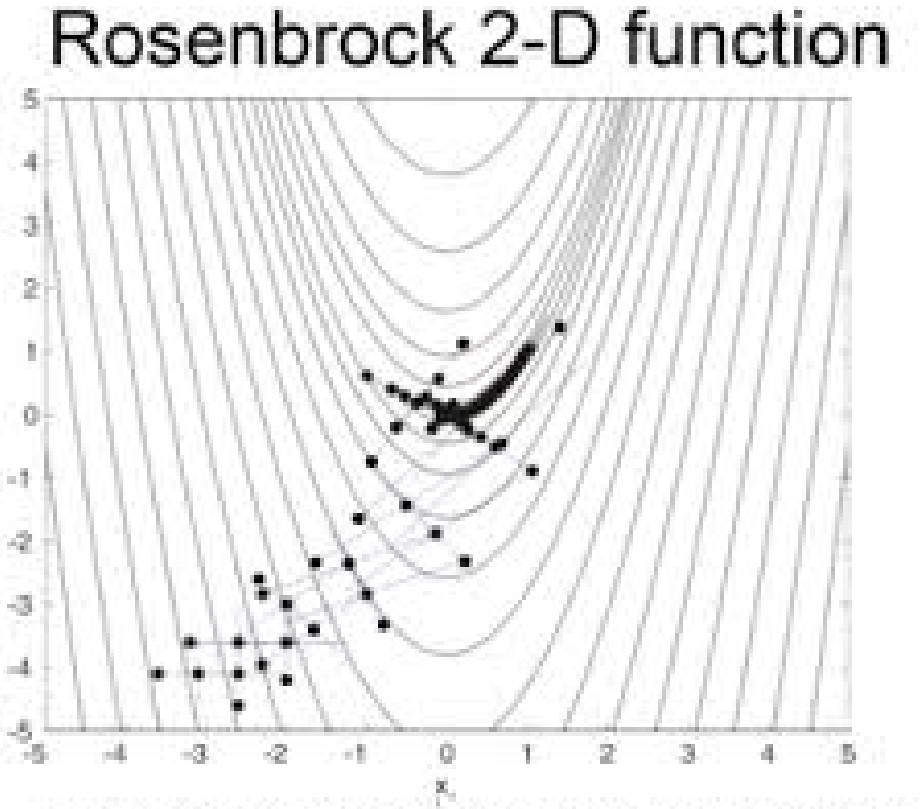
The Rosenbrock function

The Rosenbrock function is a non-convex function used as a performance test problem for optimization algorithms introduced by Howard H. Rosenbrock in 1960. It is also known as Rosenbrock's banana function. The function is defined by

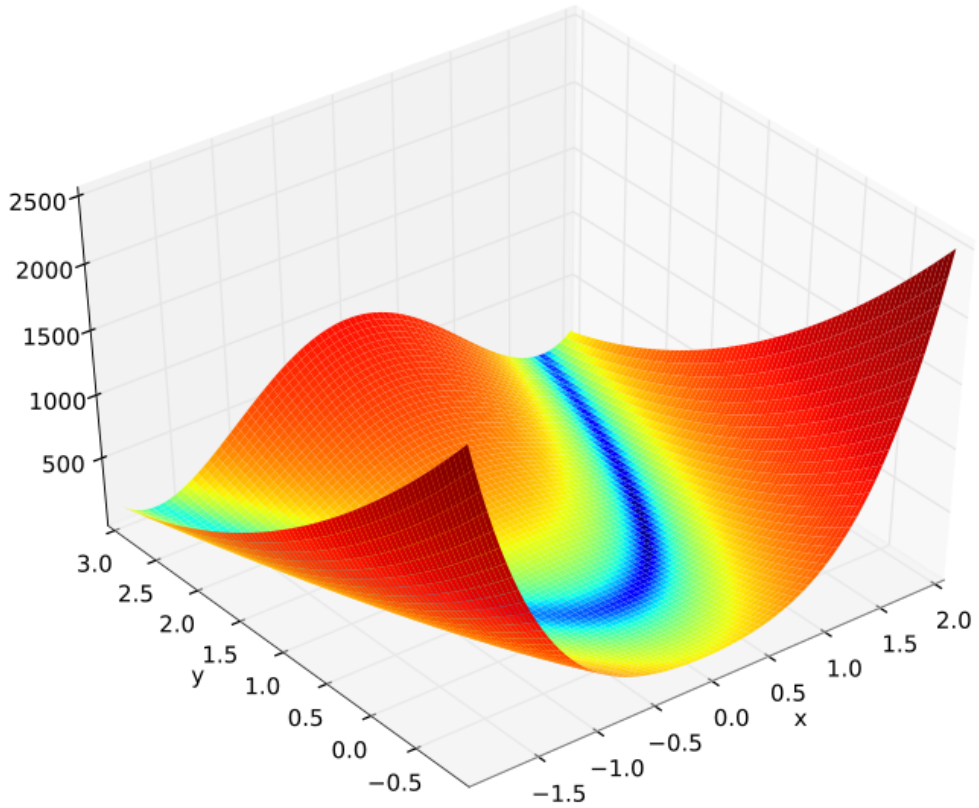
$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2.$$

It has a global minimum at $(x, y) = (1, 1)$, where $f(x, y) = 0$.

The Rosenbrock function



The Rosenbrock function



The Himmelblau function

In mathematical optimization, the Himmelblau function is a multi-modal function, used to test the performance of optimization algorithms. The function is defined by

$$f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$

it has It has one local maximum at

$$x = -0.270845, y = -0.923039$$

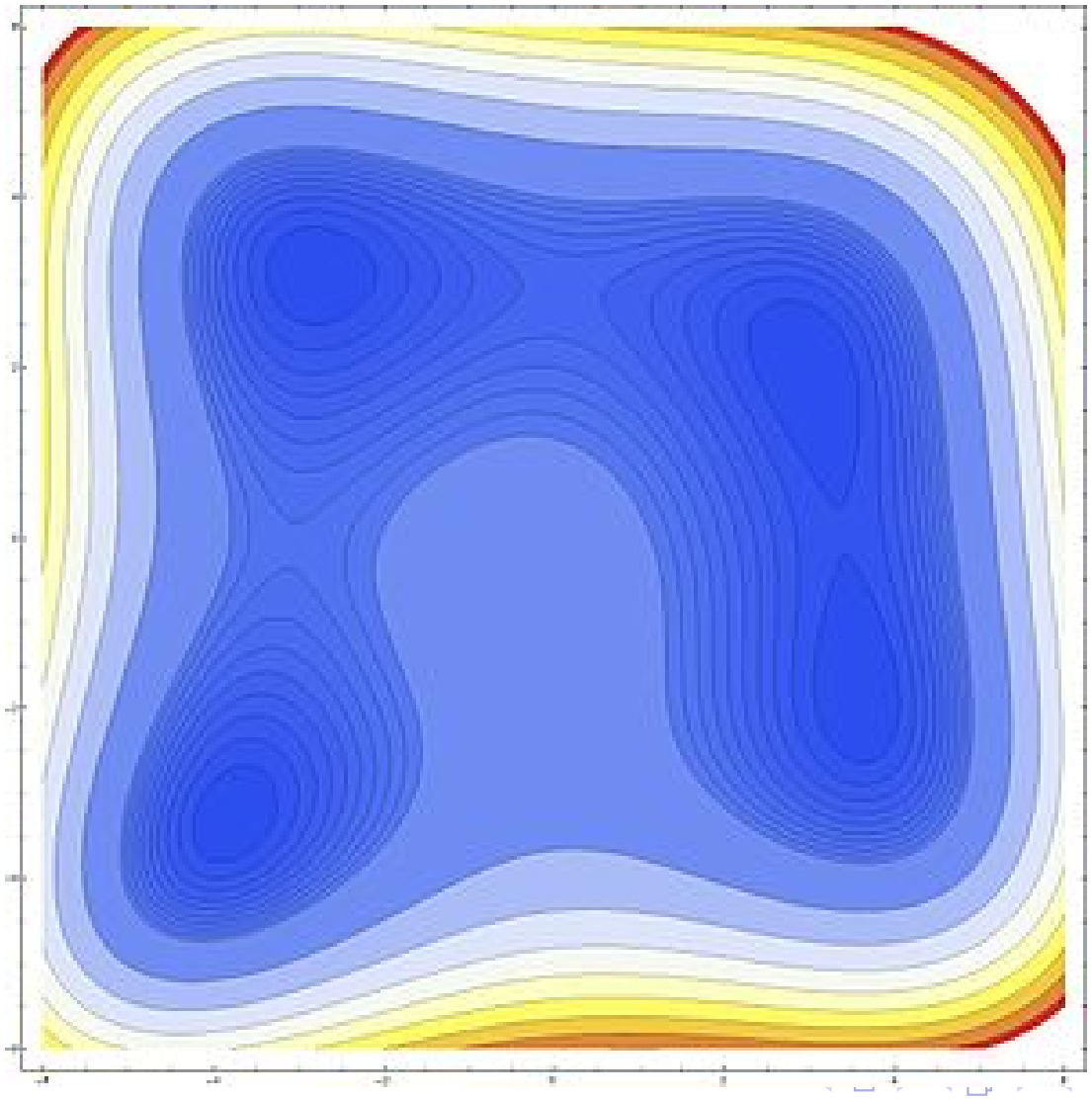
where $f(x, y) = 181.617$

The Himmelblau function

and four identical local minima

$$f(3.0, 2.0) = 0.0,$$
$$f(-2.805118, 3.131312) = 0.0,$$
$$f(-3.779310, -3.283186) = 0.0,$$
$$f(3.584428, -1.848126) = 0.0.$$

The Himmelblau function



The Himmelblau function