A method of solving a convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$.

(Presented by Academician L.V. Kantorovich 9 VII 1982)

Yu. E. Nesterov

7th April 2021.

1. The paper proposes a method for solving the convex programming problem in the Hilbert space E. Unlike most convex programming methods previously proposed, this method constructs a minimizing sequence of points $\{x_k\}_{k=0}^{\infty}$, which is not relaxational. This feature allows to minimize computational cost at each step. At the same time, for this method it is possible to obtain an unimprovable estimate of the convergence speed on the considered class of problems (see [1]).

2. Consider first the problem of unconditional minimization of the convex function $f(x)$. We will assume that the function $f(x)$ belongs to the class $C^{1,1}(E)$ i.e. there exists a constant $L > 0$, for which for all $x, y \in E$, the inequality

$$\|f'(x) - f'(y)\| \leq L\|x - y\|. \tag{1}$$

It follows from inequality (1) that for all $x, y \in E$,

$$f(y) \leq f(x) + \langle f'(x), y - x\rangle + 0.5L\|y - x\|^2. \tag{2}$$

To solve the problem $\min\{f(x) \mid x \in E\}$ with a non-empty set of minima X^*, the following method is proposed.

0) Choose the point $y_0 \in E$. Assume

$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad \alpha_{-1} = \frac{\|y_0 - z\|}{f'(y_0) - f'(z)}, \tag{3}$$

where z is any point from E, $z \neq y_0 f'(z) \neq f'(y_0)$.

1) kth iteration.

a) Calculate the smallest number $i > 0$ for which the inequality is satisfied

$$f(y_k) - f(y_k - 2^{-i}\alpha_{k-1}f'(y_k)) \geq 2^{-i-1}\alpha_{k-1}\|f'(y_k)\|^2. \tag{4}$$

b) Assume

$$\begin{gathered}\alpha_k = 2^{-i}\alpha_{k-1}, \quad x_k = y_k - \alpha_k f'(y_k), \\ a_{k+1} = \frac{1 + \sqrt{4a_k^2 + 1}}{2}, \quad y_{k+1} = x_k + \frac{(a_k - 1)(x_k - x_{k-1})}{a_{k+1}}.\end{gathered} \tag{5}$$

The method of breaking one-dimensional search (4) is similar to the method proposed in [2]. The only difference is that in (4) the step splitting at the k-th iteration is performed starting from α_{k-1} (not from one, as in [2]). Because of this (see the proof of Theorem 1), when constructing the sequence $\{x_k\}_{k=0}^{\infty}$ by the method $(3)-(5)$, no more than $O(\log_2 L)$ such fractions will be made. The recalculation of points y_k in (5) is performed using the "ravine" step. Note also that the method $(3)-(5)$ does not provide a monotonic decrease of the function $f(x)$ on sequences $\{x_k\}_{k=0}^{\infty}$, $\{y_k\}_{k=0}^{\infty}$.

Theorem 1. Let the convex function $f(x) \in C^{1,1}(E)$ and $X^* \neq \phi$. If the sequence $\{x_k\}_{k=0}^{\infty}$ is constructed by the method $(3)-(5)$, then:

1) For any $k \geq 0$,

$$f(x_k) - f^* \leq \frac{C}{(k+2)^2}, \tag{6}$$

where $C = 4L\|y_0 - x^*\|^2$, $f^* = f(x^*)$, $x^* \in X^*$.

2) To achieve accuracy ϵ in terms of the functional it is necessary:

   a) Calculate the gradient of the target function no more than $NG =]\sqrt{C/\epsilon}[$ times.

   b) Calculate the value of the function no more than $NF = 2NG +]\log_2(2L\alpha_{-1})[+1$ times.

Hereinafter $](-)[$ is the integer part of the number $(-)$.

Proof. Let $y_k(\alpha) = y_k - \alpha f'(y_k)$. From inequality (2) we obtain $f(y_k) - f(y_k(\alpha)) \geq 0.5\alpha(2 - \alpha L)\|f'(y_k)\|^2$. Consequently, as soon as $2^{-i}\alpha_{k-1}$ becomes smaller than L^{-1}, inequality (4) will be fulfilled and further α_k will not decrease. Thus $\alpha_k \geq 0.5L^{-1}$ for all $k \geq 0$.

Denote $p_k = (a_k - 1)(x_{k-1} - x_k)$. Then $p_{k+1} - x_{k+1} = p_k - x_k + a_{k+1}\alpha_{k+1}f'(y_{k+1})$. Consequently,

$$\begin{aligned}\|p_{k+1} - x_{k+1} + x^*\|^2 =& \|p_k - x_k + x^*\|^2 + 2(a_{k+1} - 1)\alpha_{k+1}\left\langle f'(y_{k+1}), p_k\right\rangle \\ &+ 2a_{k+1}\alpha_{k+1}\left\langle f'(y_{k+1}), x^* - y_{k+1}\right\rangle + a_{k+1}^2\alpha_{k+1}^2\|f'(y_{k+1})\|^2.\end{aligned}$$

Using inequality (4) and the convexity of the function $f(x)$, we obtain

$$\left\langle f'(y_{k+1}), y_{k+1} - x^*\right\rangle \geq f(x_{k+1}) - f^* + 0.5\alpha_{k+1}\|f'(y_{k+1})\|^2,$$

$$0.5\alpha_{k+1}\|f'(y_{k+1})\|^2 \leq f(y_{k+1}) - f(x_{k+1}) \leq f(x_k) - f(x_{k+1}) - a_{k+1}^{-1}\left\langle f'(y_{k+1}), p_k\right\rangle.$$

Substitute these two inequalities into the previous inequality:

$$\begin{aligned}\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 &\leq 2(a_{k+1} - 1)\alpha_{k+1}\left\langle f'(y_{k+1}), p_k\right\rangle - 2a_{k+1}\alpha_{k+1}(f(x_{k+1}) - f^*) \\ &\quad + (a_{k+1}^2 - a_{k+1})\alpha_{k+1}^2\|f'(y_{k+1})\|^2 \\ &\leq -2a_{k+1}\alpha_{k+1}(f(x_{k+1}) - f^*) \\ &\quad + 2(a_{k+1}^2 - a_{k+1})\alpha_{k+1}(f(x_k) - f(x_{k+1})) \\ &= 2\alpha_{k+1}a_k^2(f(x_k) - f^*) - 2\alpha_k a_{k+1}^2(f(x_{k+1}) - f^*) \\ &\leq 2\alpha_k a_k^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*).\end{aligned}$$

Thus,

$$
\begin{aligned}
2\alpha_k a_{k+1}^2(f(x_{k+1}) - f^*) &\le 2\alpha_{k+1} a_{k+1}^2(f(x_{k+1}) - f^*) + \|p_{k+1} - x_{k+1} + x^*\|^2 \\
&\le 2\alpha_k a_k(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2 \\
&\le 2\alpha_0 a_0^2(f(x_0) - f^*) + \|p_0 - x_0 + x^*\|^2 \\
&\le \|y_0 - x^*\|^2.
\end{aligned}
$$

It remains to to be seen that $a_{k+1} \ge a_k + 0.5 \ge 1 + 0.5(k+1)$.

From the estimate of the convergence rate (6), it follows that the number of iterations required for the method $(3) - (5)$ to achieve precision ϵ will not be greater than $]\sqrt{C/\epsilon}[-1$.

(Gtranslate. This section is shaky.) In this case, at each iteration, one gradient will be calculated and at least two objective function values. Note however, that for each additional calculation of the value of the objective function corresponds to the reduction of the value of a_k by half. Therefore, the total number of such calculations will not exceed $]\log_2(2L\alpha_{-1})[+1$.

The theorem is proved.

If the Lipschitz constant L is known for the gradient of the target function, then in method (3)-(5) it is possible to put $\alpha_k \equiv L^{-1}$ for any $k \ge 0$. In this case inequality (4) will be obviously fulfilled, and therefore the assertions of Theorem 1 remain true at $C = 2L\|y_0 - x^*\|^2$, $NG =]\|y_0 - x^*\|\sqrt{2L/\epsilon}[-1$ and $NF = 0$.

We conclude this section by showing how we can modify the method $(3) - (5)$ to solve the minimisation problem of a strongly convex function.

Suppose that for the function $f(x)$ for all $x \in E$ the inequality $f(x) - f^* \ge 0.5m\|x - x^*\|^2$, where $m > 0$ and let the constant m be known.

Let us introduce the following interrupt rule into method $(3) - (5)$:

c) We stop if

$$
k \ge 2\sqrt{2/(ma_k)} - 2. \tag{7}
$$

Let the interruption occured at the Nth step. Since in method $(3) - (5)$, $\alpha_k \ge 0.5L^{-1}$, then $N \le]4\sqrt{L/m}[-1$. At the same time

$$
f(x_N) - f^* \le \frac{2\|y_0 - x^*\|^2}{\alpha_N(N+2)^2} \le 0.25m\|y_0 - x^*\|^2 \le 0.5(f(y_0) - f^*).
$$

After the point x_N, it is necessary to update the method and start counting again by $(3) - (5)$, (7) from point x_N as from the initial point, etc.

As a result, we obtain that for every $]4\sqrt{L/m}[-1$ iterations, the residual of the function halves. Thus method $(3) - (5)$ with update (7) is unimproved (to a dimensionless constant) among first-order methods on the class of strongly convex functions from $C^{1,1}(E)$ (see [1]).

3. Consider the following extreme value problem:

$$
\min\left\{F(\overline{f}(x)) \,\middle|\, x \in Q\right\}. \tag{8}
$$

where Q is a convex closed set of E, $F(u), u \in \mathbb{R}^m$ convex throughout $\mathbb{R}^m$, positively homogeneous unit (G: degree one) function, and $\overline{f}(x) = (f_1(x), f_2(x_2), \cdots, f_m(x))$ is a vector of convex continuously differentiable functions on E. The set X^* of solutions to problem (8) is always assumed non-empty. In addition, we will always assume that the system of functions $\{F(\cdot), \overline{f}(\cdot)\}$ has the following property:

(*) If there exists a vector $\lambda \in \partial F(0)$ such that $\lambda^{(k)} < 0$, then $f_k(x)$ is a linear function.

The subdifferential of the function $F(u)$ at 0 is denoted by $\partial F(0)$ in $(*)$.

It is known for convex positively homogenoeous degree one functions that the following identity is true $F(u) \equiv \max\{\langle \lambda, u\rangle \,|\lambda \in \partial F(0)\}$. Therefore, from assumption $(*)$ follows the convexity of the function $F(\overline{f}(x))$ by E.

The problem (8) can be written in the minimax form:

$$\min\left\{\max\left\{\left\langle \lambda, \overline{f}(x)\right\rangle \,\middle|\, \lambda \in \partial F(0)\right\} \,\middle|\, x \in Q\right\}. \tag{9}$$

It can be shown that the non-emptiness of the set X^* and assumption $(*)$ imply the existence of a saddle point (λ^*, x^*) in (9). Therefore the set of saddle points of problem (9) can be represented as $\Omega^* = \Lambda^* \times X^*$, where $\Lambda^* = \text{argmax}\{\Psi(\lambda) \mid \lambda \in \partial F(0)\}$ and $\Psi(\lambda) = \min\{\langle\lambda, f(x)\rangle \mid x \in Q\}$. The task is

$$\max\left\{\Psi(\lambda) \,|\, \lambda \in \partial F(0) \cap \text{dom}\Psi(\cdot)\right\}.$$

We will call the problem dual to (8).

Let in problem (8) the functions $f_k(x)$, $k = 1, \cdots, m$, belong to the class $C^{1,1}(E)$ with constants $L^{(k)} \geq 0$. Denote $\overline{L} = (L^{(1)}, L^{(2)}, \cdots, L^{(m)})$.

Consider the function

$$\Phi(y, A, z) = F(\overline{f}(y, z)) + 0.5A\|y - z\|^2,$$

where

$$\overline{f}(y, z) = (f^{(1)}(y, z), f^{(2)}(y, z), \ldots, f^{(m)}(y, z)),$$

$$f^{(k)}(y, z) = f_k(y) + \langle f'(y), z - y\rangle\,, \quad k = 1, \ldots, m,$$

and where A is a postive constant. Denote

$$\Phi^*(y, A) = \min\left\{\Phi(y, A, z) \,|\, z \in Q\right\}, \quad T(y, A) = \text{argmin}\left\{\Phi(y, A, z) \,|\, z \in Q\right\}.$$

Note that the mapping $y \to T(y, A)$ is a natural generalization for problem (8) the “gradient” mapping introduced in [1] in connection with the investigation of methods for minimization of functions of the form $\max_{1 \leq k \leq m} f_k(x)$. For the mapping $y \to T(y, A)$ (as for the “gradient” mapping from [1]), for all $x \in Q$, $y \in E$, $A \geq 0$, the inequality

$$\Phi^*(y, A) + A\,\langle y - T(y, A), x - y\rangle + 0.5A\|y - T(y, A)\|^2 \leq F(\overline{f}(x)), \tag{10}$$

and if $A > F(L)$, then

$$\Phi^*(y, A) \geq F(\overline{f}(T(y, A))).$$

To solve the problem (8) the following method is proposed.

0) Choose the point $y_0 \in E$. Assume

$$k = 0, \quad \alpha_0 = 1, \quad x_{-1} = y_0, \quad A_{-1} = F(\overline{L}_0), \tag{11}$$

where $\overline{L}_0 = (L_0^{(1)}, L_0^{(2)}, \cdots, L_0^{(m)})$, $L_0^{(k)} = \|f_k'(y_0) - f_k'(z)\|/\|y_0 - z\|$, and z is a random (G: arbitrary) point from E, $z \neq y_0$.

1) kth iteration.

a) We assign the smallest number $i \geq 0$, for which the inequality is satisfied

$$\Phi^*(y_k, 2^i A_{k-1}) \geq F(\overline{f}(T(y_k, 2^i A_{k-1}))). \tag{12}$$

b) Assume

$$\begin{aligned} &A_k = 2^i A_{k-1}, \quad x_k = T(y_k, A_k), \\ &a_{k+1} = \frac{(1 + \sqrt{4a_k^2 + 1})}{2}, \quad y_{k+1} = x_k + \frac{(a_k - 1)(x_k - x_{k-1})}{a_{k+1}}. \end{aligned} \tag{13}$$

It is easy to see that method (3) – (5) is simply another form of the method (11) – (13) for the unconstrained minimisation problem (i.e. when in (8) $m = 1$, $F(y) = y$, $Q = E$).

Theorem 2. If the sequence $\{x_k\}_{k=0}^{\infty}$ is constructed by method (11) – (13), then

1) For any $k \geq 0$,

$$F(\overline{f}(x_k)) - F(\overline{f}(x^*)) \leq \frac{C_1}{(k+2)^2},$$

where $C_1 = 4f(\overline{L})\|y_0 - x^*\|^2, x^* \in X^*$.

2) To achieve accuracy ϵ by function it is necessary to:

a) Solve an auxiliary problem $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ no more than $]\sqrt{C_1/\epsilon}[+] \max \log_2(F(\overline{L})/A_{-1}), 0[$ times.

b) Calculate the set of gradients $f_1'(y), f_2'(y), \cdots, f_m'(y)$ no more than $]\sqrt{C_1/\epsilon}[$ times.

c) Calculate the vector function $\overline{f}(x)$ no more than $2]\sqrt{C_1/\epsilon}[+] \max\{\log_2(F(\overline{L})/A_{-1}, 0\}[$ times.

Theorem 2 is proved in much the same way as Theorem 1. It it is necessary only use inequality (10) instead of inequality (2), and the analogue of vector $\alpha_k f'(y_k)$ will be vector $y_k - T(y_k, A_k)$, and the analogue of α_k will be A_k^{-1}.

Just as in method (3) – (5), in method (11) – (13) we can take into account information about the constant $F(\overline{L})$ and the strong convexity parameter of the function $F(\overline{f}(x)) - m$ (for this however, it is necessary that $y_0 \in Q$).

In conclusion, we note two important special cases of problem (8), in which the auxiliary problem $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ turns out to be quite simple.

a) Minimization of a smooth convex function on a simple set. By a simple set we mean a set for which the design (G: projection) operator can be written explicitly. In this case the problem (8) $m = 1, F(y) = y$\$ and in the method (11) – (13)

$$\Phi^*(y, A) = f(y) - 0.5A^{-1}\|f'(y)\|^2 + 0.5A\|T(y, A) - y + A^{-1}f'(y)\|^2,$$

where $T(y, A) = \operatorname{argmin}\left\{\|y - A^{-1}f^{-1}(y) - z\| \,\middle|\, z \in Q\right\}$.

b) Unconditional minimization (in problem (8) $Q \equiv E$). In this case the the auxiliary problem $\min\{\Phi(y, A, x) \mid x \in E\}$ is equivalent to the following dual problem:

$$\max\left\{-0.5A^{-1}\left\|\sum_{k=1}^{m}\lambda^{(k)}f_k'(y)\right\|^2+\sum_{k=1}^{m}\lambda^{(k)}f_k(y)\;\middle|\;(\lambda^{(1)},\lambda^{(2)},\ldots,\lambda^{(m)})\in\partial F(0)\right\}. \tag{14}$$

In this case, $T(y,A)=y-A^{-1}\sum_{k=1}^{m}\lambda^{(k)}(y)f_k'(y)$, where $\lambda^{(k)}(y), k=1,2,\ldots,m$ are solutions of problem (14) at a fixed $y\in E$. Note that the set $\partial F(0)$ is usually given by simple constraints either linear or quadratic. In such cases the problem (14) is a standard quadratic programming problem.

The author is sincerely grateful to A. S. Nemirovsky for his conversations, which stimulated his interest in the issues discussed.

Central Institute of the Economcis and Mathematics Received Academy of Sciences of the USSR, Moscow.

Retrieved from 19 VII 1982

Literature

1. Nemirovsky A.S., Yudin D.B. Complexity of Problems and Efficiency of Methods of Optimization. Moscow: Nauka, 1979.
2. Pshenichny B.N., Danilin YM. Numerical methods in extreme problems. Moscow: Nauka, 1975.