Докл. Акад. Наук СССР
Том 269 (1983), № 3

Soviet Math. Dokl.
Vol. 27 (1983), No. 2

A METHOD OF SOLVING A CONVEX PROGRAMMING PROBLEM WITH CONVERGENCE RATE $O(1/k^2)$

UDC 51

YU. E. NESTEROV

1. In this note we propose a method of solving a convex programming problem in a Hilbert space E. Unlike the majority of convex programming methods proposed earlier, this method constructs a minimizing sequence of points $\{x_k\}_0^\infty$ that is not relaxational. This property allows us to reduce the amount of computation at each step to a minimum. At the same time, it is possible to obtain an estimate of convergence rate that cannot be improved for the class of problems under consideration (see [1]).

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

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

f(x) is L smooth
f(x) is continuous and differentiable.

From (1) it follows that for all $x, y \in E$

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

f(x) is convex

To solve the problem $\min\{f(x) \mid x \in E\}$ with a nonempty set X^* of minima we propose the following method.

0) Select a point $y_0 \in E$. Put

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

$k = 0$
$a_0 = 1$
$x_{-1} = y_0$
$\alpha_{-1} = \dfrac{\|y_0 - z\|}{\|\nabla f(y_0) - \nabla f(z_0)\|}$

where z is an arbitrary point in E, $z \neq y_0$ and $f'(z) \neq f'(y_0)$.

1) *kth iteration*. a) Calculate the smallest index $i \geqslant 0$ for which

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

k'th iteration

b) Put

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

update variables

The way in which the one-dimensional search (4) is halted is similar to that proposed in [2]. The difference is only that in (4) the subdivision in the kth iteration is done starting with α_{k-1} (and not with 1 as in [2]). In view of this (see the proof of Theorem 1), when the sequence $\{x_k\}_0^\infty$ is constructed by method (3)–(5), no more than $\boxed{O(\log_2 L)}$ such subdivisions will be made. The recalculation of the points y_k in (5) is done using a "ravine" step.

runtime of the algo?

1980 *Mathematics Subject Classification*. Primary 90C25.

Let us also remark that method (3)–(5) does not guarantee a monotone decrease of $f(x)$ on the sequences $\{x_k\}_0^\infty$ and $\{y_k\}_0^\infty$.

optimum set is nonempty

THEOREM 1. *Let $f(x)$ be a convex function in $C^{1,1}(E)$, and suppose $X^* \neq \varnothing$. If the sequence $\{x_k\}_0^\infty$ is constructed by method (3)–(5), then the following assertions are true:*

1) *For any $k \geqslant 0$;*

$$f(x_k) - f(\vec{x}^*) \le \frac{C}{(k+2)^2}$$

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

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

2) *In order to achieve accuracy ε with respect to the functional, one needs*

$NG = \lfloor \sqrt{C/\varepsilon} \rfloor$

a) *to compute the gradient of the objective function no more than $NG =]\sqrt{C/\varepsilon}[$ times, and*

b) *to evaluate the objective function no more than $NF = 2NG +]\log_2(2L\alpha_{-1})[+ 1$ times.*

$NF = 2(NG) + \lfloor \log_2(2L\alpha_{-1}) \rfloor + 1$

Here and in what follows, $](\cdot)[$ is the integer part of the number $(\cdot)$.

PROOF. Let $y_k(\alpha) = y_k - \alpha f'(y_k)$. From (2) we obtain

$$f(y_k) - f(y_k(\alpha)) \geqslant 0.5\alpha(2 - \alpha L)\|f'(y_k)\|^2.$$

Consequently, as soon as $2^{-i}\alpha_{k-1}$ becomes less than L^{-1}, inequality (4) will be satisfied and α_k will not be further decreased. Thus $\boxed{\alpha_k \geqslant 0.5L^{-1} \text{ for all } k \geqslant 0.}$

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

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

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

$$\langle f'(y_{k+1}), y_{k+1} - x^* \rangle \geqslant f(x_{k+1}) - f^* + 0.5\alpha_{k+1}\|f'(y_{k+1})\|^2,$$
$$0.5\alpha_{k+1}\|f'(y_{k+1})\|^2 \leqslant f(y_{k+1}) - f(x_{k+1}) \leqslant f(x_k) - f(x_{k+1})$$
$$- a_{k+1}^{-1}\langle f'(y_{k+1}), p_k \rangle.$$

We substitute these two inequalities into the preceding equality:

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

Thus

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

It remains to observe that $a_{k+1} \geqslant a_k + 0.5 \geqslant 1 + 0.5(k+1)$.

It follows from the estimate of the convergence rate (6) that the number of iterations method (3)–(5) needs to achieve accuracy ε will be no greater than $]\sqrt{C/\varepsilon}[- 1$. During each iteration, one gradient and at least two values of the objective function will have to

$\lfloor \log_2(2L\alpha_{-1}) \rfloor + 1$ evaluations of objective.

be calculated. Let us remark, however, that to each additional evaluation of the objective function corresponds a halving of α_k. Therefore the total number of such evaluations will not exceed $]\log_2(2L\alpha_{-1})[\, + 1$. This completes the proof of the theorem. bruh.

If the Lipschitz constant L is known for the gradient of the objective function, then one can take $\alpha_k \equiv L^{-1}$ in the method (3)–(5) for any $k \geqslant 0$. In this case inequality (4) is certain to hold, and therefore Theorem 1 remains valid for $C = 2L\|y_0 - x^*\|^2$, $Ng =]\|y_0 - x^*\|\sqrt{2L/\varepsilon}[\, - 1$ and $NF = 0$.

$\alpha_k \equiv \frac{1}{L}$ for any $k \geq 0$.

To conclude this section we will show how one may modify the method (3)–(5) to solve the problem of minimizing a strictly convex function.

Assume that $f(x) - f^* \geqslant 0.5m\|x - x^*\|^2$ for all $x \in E$, where $m > 0$, and suppose the constant m is known. ↳ This comes from strong convexity!

We introduce the following halting rule in the method (3)–(5).

c) We stop when

$$k \geqslant 2\sqrt{2/(m\alpha_k)} - 2. \tag{7}$$

Suppose that the halting has occurred in the Nth step. Since $\alpha_k \geqslant 0.5L^{-1}$ in the method (3)–(5), one has $N \leqslant]4\sqrt{L/m}[\, - 1$. At the same time, $N \leq \lfloor 4\sqrt{L/m} \rfloor - 1$

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

After the point x_N has been obtained, it is necessary to restart the method and again begin calculating, by the method (3)–(5), (7), from the point x_N as the initial point, etc.

As a result we obtain that after each $]4\sqrt{L/m}[\, - 1$ iterations the residual with respect to the function decreases by a factor of 2. Thus the method (3)–(5) with renewal (7) cannot be improved (up to a dimensionless constant) among methods of first order on the class of strictly convex functions in $C^{1,1}(E)$ (see [1]).

halves the loss after that many iterations

3. Consider the following extremal problem:

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

where Q is a convex closed set in E, $F(u)$, with $u \in R^m$, is a function convex on all of R^m, positive homogeneous of degree one, and $\bar{f}(x) = (f_1(x), \ldots, f_m(x))$ is a vector of convex continuously differentiable functions on E. The set X^* of solutions of (8) is always assumed to be nonempty. In addition to this, we will always assume that the system of functions $\{F(\cdot), \bar{f}(\cdot)\}$ 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 notation $\partial F(0)$ means the subdifferential of the function $F(u)$ at 0.

As is well known, the identity $F(u) \equiv \max\{\langle \lambda, u\rangle | \lambda \in \partial F(0)\}$ holds for convex functions that are positive homogeneous of degree one. Therefore the assumption ($*$) implies the convexity of the function $F(\bar{f}(x))$ on all of E.

Problem (8) can be written in minimax form:

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

One can show that the fact that the set X^* is nonempty and the assumption ($*$) imply the existence of a saddle point (λ^*, x^*) for problem (9). Therefore the set of saddle points of problem (9) can be written as $\Omega^* = \Lambda^* \times X^*$, where

$$\Lambda^* = \operatorname{Arg\,max}\{\Psi(\lambda) | \lambda \in \partial F(0)\}, \qquad \Psi(\lambda) = \min\{\langle \lambda, f(x)\rangle | x \in Q\}.$$

The problem

$$\max\{\Psi(\lambda) \mid \lambda \in \partial F(0) \cap \operatorname{dom}\Psi(\cdot)\}$$

will be called the problem dual to (8).

Suppose the functions $f_k(x)$, $k = 1,\ldots,m$, in problem (8) belong to the class $C^{1,1}(E)$ with constants $L^{(k)} \geqslant 0$. Let $\bar{L} = (L^{(1)},\ldots,L^{(m)})$.

Consider the function

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

where

$$\bar{f}(y, z) = (f^{(1)}(y, z),\ldots,f^{(m)}(y, x)),$$
$$f^{(k)}(y, z) = f_k(y) + \langle f'(y), z - y\rangle, \qquad k = 1, 2,\ldots,m,$$

and A is a positive constant. Let

$$\Phi^*(y, A) = \min\{\Phi(y, A, z) \mid z \in Q\}, \qquad T(y, A) = \arg\min\{\Phi(y, A, z) \mid z \in Q\}.$$

Observe that the mapping $y \to T(y, a)$ is a natural generalization, for problem (8), of the "gradient" mapping introduced in [1] in connection with the investigation of methods of minimizing functions of the form $\max_{1 \leqslant k \leqslant m} f_k(x)$. For the mapping $y \to T(y, A)$ (as well as for the "gradient" mapping of [1]) we have

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

for all $x \in Q$, $y \in E$ and $A \geqslant 0$, and if $A \geqslant F(L)$, then

$$\Phi^*(y, A) \geqslant F(\bar{f}(T(v, A))).$$

To solve problem (8) we propose the following method.

0) Select a point $y_0 \in E$. Put

$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad A_{-1} = F(\bar{L}_0), \tag{11}$$

where $\bar{L}_0 = (L_0^{(1)},\ldots,L_0^{(m)})$, $L_0^{(k)} = \|f_k'(y_0) - f_k'(z)\|/\|y_0 - z\|$ and z is an arbitrary point in E, $z \neq y_0$.

1) *kth iteration.* a) Calculate the smallest index $i \geqslant 0$ for which

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

b) Put $A_k = 2^i A_{k-1}$, $x_k = T(y_k, A_k)$ and

$$\begin{aligned} a_{k+1} &= \left(1 + \sqrt{4a_k^2 + 1}\right)/2, \\ y_{k+1} &= x_k + (a_k - 1)(x_k - x_{k-1})/a_{k+1}. \end{aligned} \tag{13}$$

It is not hard to see that the method (3)–(5) is simply another form of writing the method (11)–(13) for the unconstrained minimization problem (i.e., when $m = 1$, $F(y) = y$ and $Q = E$ in (8)).

THEOREM 2. *If the sequence $\{x_k\}_0^\infty$ is constructed by method* (11)–(13), *then the following assertions are true*:

1) *For any $k \geqslant 0$*

$$F(\bar{f}(x_k)) - F(\bar{f}(x^*)) \leqslant C_1/(k + 2)^2,$$

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

2) *To obtain accuracy* ε *with respect to the functional, one needs*

a) *to solve an auxiliary problem* $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ *no more than*

$$]\sqrt{C_1/\varepsilon}[\ +]\max\{\log_2(F(\bar{L})/A_{-1}), 0\}[$$

times,

b) *to evaluate the collection of gradients* $f_1'(y), \ldots, f_m'(y)$ *no more than* $]\sqrt{C_1/\varepsilon}[$ *times, and*

c) *to evaluate the vector-valued function* $\bar{f}(x)$ *at most*

$$2]\sqrt{C_1/\varepsilon}[\ +]\max\{\log_2(F(\bar{L})/A_{-1}), 0\}[$$

times.

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

Just as in the method (3)–(5), in the method (11)–(13) one can take into account information about the constant $F(\bar{L})$ and the parameter of strict convexity of the function $F(\bar{f}(x)) - m$ (for this, of course, we must have $y_0 \in Q$).

In conclusion let us mention 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 rather simple.

a) *Minimization of a smooth function on a simple set*. By a simple set we understand a set for which the projection operator can be written in explicit form. In this case $m = 1$ and $F(y) = y$ in problem (8), and

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

in the method (11)–(13), where

$$T(y, A) = \arg\min\{\|y - A^{-1}f'(y) - z\| \mid z \in Q\}.$$

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

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

Here

$$T(y, A) = y - A^{-1}\sum_{k=1}^{m} \lambda^{(k)}(y) f_k'(y),$$

where the $\lambda^{(k)}(y)$, $k = 1, \ldots, m$, are solutions of problem (14) for fixed $y \in E$. Let us remark that the set $\partial F(0)$ is usually given by simple constraints—linear or quadratic. In such cases problem (14) is the standard quadratic programming problem.

The author expresses his sincere appreciation to A. S. Nemirovskiĭ for discussions that stimulated his interest in the questions considered here.

Central Economico-Mathematical Institute
Academy of Sciences of the USSR

Received 19/JULY/82

BIBLIOGRAPHY

1. A. S. Nemirovskiĭ and D. B. Yudin, *Complexity of problems and efficiency of optimization methods*, "Nauka" Moscow, 1979. (Russian)

2. B. N. Pshenichnyĭ and Yu. M. Danilin, *Numerical methods in extremal problems*, "Nauka", Moscow, 1975; French transl., "Mir", Moscow, 1977.

Translated by A. ROSA