Working Paper

Convex Optimization by Radial Search

Yuri M. Ermoliev
Andrzej Ruszczyński

WP-95-036
April 1995

IIASA
International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Convex Optimization by Radial Search

Yuri M. Ermoliev
Andrzej Ruszczyński

WP-95-036
April 1995

IIASA International Institute for Applied Systems Analysis ◻ A-2361 Laxenburg ◻ Austria
Telephone: +43 2236 807 ◻ Fax: +43 2236 71313 ◻ E-Mail: info@iiasa.ac.at

Abstract

A convex nonsmooth optimization problem is replaced by a sequence of line search problems along recursively updated rays. Convergence of the method is proved and applications to linear inequalities, constraint aggregation and saddle point seeking indicated.

Key words: Nonsmooth optimization, subgradient methods, aggregation.

Convex Optimization by Radial Search

Yuri M. Ermoliev
Andrzej Ruszczyński

1 The method

The objective of this note is to present a new algorithmic concept for convex optimization problems of the form:

$$\min f(x), \quad x \in \mathbb{R}^n. \tag{1.1}$$

We assume that the function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ satisfies the following assumptions:

(A1) f is convex, closed and co-finite, i.e. $\sup_x\{\langle y, x\rangle - f(x)\} < \infty$ for all $y \in \mathbb{R}^n$;

(A2) $0 \in \operatorname{int} \operatorname{dom} f$.

Consider the following method.

ALGORITHM 1

Step 0: Choose $s^0 \in \mathbb{R}^n$ and $\sigma \in (0, 1)$; set $k = 0$.

Step 1: Find $x^k = -\mu_k s^k$ by minimizing f along the ray $\{-\mu s^k : \mu \geq 0\}$.

Step 2: Find a subgradient $g^k \in \partial f(x^k)$ such that $\left|\langle s^k, g^k\rangle\right| \leq \sigma|s^k|^2$ if $x^k \neq 0$ and $\langle s^k, g^k\rangle \leq \sigma|s^k|^2$ if $x^k = 0$.

Step 3: Set $s^{k+1} = (1 - \tau_k)s^k + \tau_k g^k$, increase k by one and go to Step 1.

Our method employs line search, as some of the bundle methods of [3, 4], but has a simple direction-generating rule, close to the subgradient averaging employed in some stochastic subgradient algorithms [1, 6]. Moreover, we do not increment x^k in successive directions, but we stay at one point (here 0) and we explore the space along selected rays. The method emerged from our recent work [2] on constraint aggregation schemes.

Throughout the paper we shall assume the following conditions on the stepsizes $\{\tau_k\}$.

(A3) $\tau_k \in [0, 1]$;

(A4) $\tau_k \to 0$;

(A5) $\sum_{k=0}^{\infty} \tau_k = \infty$.

We shall base our analysis on the following lemma (see [2]).

Lemma 1.1. *Let the sequences* $\{\beta_k\}$, $\{\tau_k\}$, $\{\delta_k\}$ *and* $\{\gamma_k\}$ *satisfy the inequality*

$$0 \le \beta_{k+1} \le \beta_k - \tau_k \delta_k + \gamma_k. \tag{1.2}$$

If

(i) $\liminf \delta_k \ge 0$;

(ii) *for every subsequence* $\{k_i\} \subset \mathbb{N}$ *one has* $[\liminf \beta_{k_i} > 0] \Rightarrow [\liminf \delta_{k_i} > 0]$;

(iii) $\tau_k \ge 0,\ \lim \tau_k = 0,\ \sum_{k=0}^{\infty} \tau_k = \infty$;

(iv) $\lim \gamma_k / \tau_k = 0$,

then $\lim_{k\to\infty} \beta_k = 0$.

Proof. Suppose that $\liminf \delta_k = \delta > 0$. Then (1.2) for large k yields $\beta_{k+1} \le \beta_k - \tau_k \delta/2 + \gamma_k \le \beta_k - \tau_k \delta/4$. This contradicts (iii). Therefore $\liminf \delta_k = 0$. By (ii) there is a subsequence $\{k_i\}$ such that $\beta_{k_i} \to 0$. Suppose that there is another subsequence $\{s_j\}$ such that $\beta_{s_j} \ge \beta > 0$ for $j = 0, 1, 2, \ldots$. With no loss of generality we may assume that $k_1 < s_1 < k_2 < s_2 \ldots$. By (i), (iii) and (iv), for all sufficiently large j there must exist indices $r_j \in [k_j, s_j]$ such that $\beta_{r_j} > \beta/2$ and $\beta_{r_{j+1}} > \beta_{r_j}$. But then, by (ii), $\liminf \delta_{r_j} = \delta > 0$ and we obtain a contradiction with (1.2) for large j. □

Lemma 1.2. *There exists a constant* C *such that for all* k *one has* $|g^k| \le C(1 + |s^k|)$.

Proof. Denote $f_{\min} = \min f(x)$. By (A2), $f_{\min} > -\infty$. For every $\epsilon > 0$ we have

$$\begin{aligned} f\left(\frac{\epsilon g^k}{|g^k|}\right) &\ge f(x^k) + \left\langle \frac{\epsilon g^k}{|g^k|} + x^k, g^k \right\rangle \\ &\ge f_{\min} + \epsilon |g^k| - \mu_k \langle s^k, g^k \rangle. \end{aligned}$$

Using the conditions of Step 2 we obtain

$$f\left(\frac{\epsilon g^k}{|g^k|}\right) \ge f_{\min} + \epsilon |g^k| - \sigma \mu_k |s^k|^2.$$

By (A2), the set $X_0 = \{x \in \mathbb{R}^n : \ f(x) \le f(0)\}$ has a finite diameter d. Therefore $\mu_k |s^k| \le d$. Moreover, f is finite around 0, so for some small but fixed $\epsilon > 0$ and some C_1, $f(\epsilon g^k / |g^k|) \le C_1$ for all k. The last inequality then implies that

$$\epsilon |g^k| \le C_1 - f_{\min} + \sigma d |s^k|,$$

which yields the required result. □

Lemma 1.3. $\lim\limits_{k\to\infty} s^k = 0$.

Proof. By the conditions of Step 2,

$$
\begin{aligned}
|s^{k+1}|^2 &= (1-\tau_k)^2|s^k|^2 + 2\tau_k(1-\tau_k)\langle s^k, g^k\rangle + \tau_k^2|g^k|^2 \\
&\le (1-2(1-\sigma)\tau_k + \tau_k^2)|s^k|^2 + \tau_k^2|g^k|^2.
\end{aligned} \tag{1.3}
$$

By Lemma 1.1,

$$|g^k|^2 \le C^2(1+|s^k|)^2 \le 2C^2(1+|s^k|^2).$$

Therefore,

$$|s^{k+1}|^2 \le (1-2(1-\sigma)\tau_k + (2C^2+1)\tau_k^2)|s^k|^2 + 2C^2\tau_k^2.$$

By (A4), for all sufficiently large k one has $\tau_k \le (1-\sigma)/(2C^2+1)$, so

$$|s^{k+1}|^2 \le (1-(1-\sigma)\tau_k)|s^k|^2 + 2C^2\tau_k^2.$$

The required result follows now from Lemma 1.1. □

Theorem 1.4. *Assume (A1)-(A5). Then for the sequence $\{x^k\}$ generated by Algorithm 1 one has*

$$\liminf f(x^k) = \min_{x\in\mathbb{R}^n} f(x).$$

Proof. Consider the conjugate function $f^*(\cdot) = \max_x\{\langle x,\cdot\rangle - f(x)\}$ (see, e.g., [3, 5]). It is convex and (by assumption) finite everywhere. From the convexity of f^* we get

$$f^*(s^{k+1}) \le (1-\tau_k)f^*(s^k) + \tau_k f^*(g^k).$$

From Fenchel's equality (see, e.g. [5, Thm. 23.5]) and conditions of Step 2 we obtain

$$
\begin{aligned}
f^*(g^k) &= -f(x^k) + \langle x^k, g^k\rangle \\
&= -f(x^k) - \mu_k\langle s^k, g^k\rangle \\
&\le -f(x^k) + \mu_k\sigma|s^k|^2 \\
&\le -f(x^k) + \sigma d|s^k|,
\end{aligned}
$$

where d is the upper bound on $|x^k| = \mu_k|s^k|$. Combining the last two inequalities we obtain

$$f^*(s^{k+1}) \le f^*(s^k) - \tau_k(f^*(s^k) + f(x^k) - \sigma d|s^k|). \tag{1.4}$$

By the continuity of f^*, $f^*(s^k) \to f^*(0) = -f_{\min}$. Suppose that $f(x^k) \ge f_{\min} + \epsilon$ for all k, where $\epsilon > 0$. Then (1.4), Lemma 1.3 and (A5) imply that $f^*(s^k) \to -\infty$, a contradiction. Therefore $\liminf f(x^k) = f_{\min}$. □

A stronger result can be obtained for the sequence of averages.

Theorem 1.5. *Let the assumptions of Theorem 1.4 be satisfied. Then for the sequence of averages*

$$\bar{x}^{k+1} = (1-\tau_k)\bar{x}^k + \tau_k x^k, \quad k = 0,1,2,\ldots,$$

where $\{x^k\}$ is generated by Algorithm 1, one has

$$\lim_{k\to\infty} f(\bar{x}^k) = \min_{x\in\mathbb{R}^n} f(x).$$

Proof. From the convexity of f and f^* we obtain

$$f(\bar{x}^{k+1}) \leq (1-\tau_k)f(\bar{x}^k) + \tau_k f(x^k),$$

$$f^*(s^{k+1}) \leq (1-\tau_k)f^*(s^k) + \tau_k f^*(g^k).$$

Adding both sides yields

$$f(\bar{x}^{k+1}) + f^*(s^{k+1}) \leq (1-\tau_k)(f(\bar{x}^k) + f^*(s^k)) + \tau_k\langle x^k, g^k\rangle.$$

because $f(x^k) + f^*(g^k) = \langle x^k, g^k\rangle$ [5, Thm. 23.5]. By the conditions of Step 2, $\langle x^k, g^k\rangle \leq \mu_k\sigma|s^k|^2 \leq d|s^k|$, where d is the upper bound on $|x^k|$. Therefore,

$$\max(0, f(\bar{x}^{k+1}) + f^*(s^{k+1})) \leq (1-\tau_k)\max(0, f(\bar{x}^k) + f^*(s^k)) + \tau_k d|s^k|.$$

Since $|s^k| \to 0$ by Lemma 1.3, using Lemma 1.1 we conclude that

$$\lim_{k\to\infty} \max(0, f(\bar{x}^k) + f^*(s^k)) = 0. \tag{1.5}$$

With $f^*(s^k) \to f^*(0) = -f_{\min}$, the required result follows from (1.5). □

2 Explicit non-negativity constraints

The concept introduced in section 1 applies, of course, to constrained problems, because we allow $+\infty$ as the value of f. For example, simple inequalities $x \geq 0$ can be dealt with by moving the center 0 to some $\tilde{x} > 0$. It is, however, more convenient to treat them explicitly.

Consider the problem

$$\min_{x\geq 0} f(x)$$

under the same assumptions as before. Then we can still apply the method described in section 1, with the following modifications.

ALGORITHM 2

Step 0: Choose $s^0 \in \mathbb{R}^n$ and $\sigma \in (0,1)$; set $k = 0$.

Step 1: Find $x^k = \mu_k d^k$ by minimizing f along the ray $\{\mu d^k : \mu \geq 0\}$, where d^k is the projection of $-s^k$ onto the positive orthant: $d_j^k = \max(0, -s_j^k)$, $j = 1, \ldots, n$.

Step 2: Find a subgradient $g^k \in \partial f(x^k)$ such that $\left|\langle d^k, g^k\rangle\right| \leq \sigma|d^k|^2$ if $x^k \neq 0$ and $\langle d^k, g^k\rangle \geq -\sigma|d^k|^2$ if $x^k = 0$.

Step 3: Set $s^{k+1} = (1-\tau_k)s^k + \tau_k g^k$, with $\tau_k \in [0,1]$, increase k by one and go to Step 1.

The convergence properties remain unchanged.

Theorem 2.1. *Let the assumptions of Theorem 1.4 be satisfied. Then for the sequence* $\{x^k\}$ *generated by Algorithm 2 one has*

$$\liminf f(x^k) = \min_{x \geq 0} f(x).$$

Proof. We shall derive a counterpart of the key inequality (1.3). From the definition of d^k one obtains

$$-s^{k+1} \leq (1 - \tau_k)d^k - \tau_k g^k.$$

In the above vector inequality, for the components j such that $-s_j^{k+1} > 0$ the absolute value of the right hand side is not less than $|s_j^{k+1}|$, so

$$\begin{aligned} |d^{k+1}|^2 &\leq |(1 - \tau_k)d^k - \tau_k g^k|^2 \\ &= (1 - \tau_k)^2 |d^k|^2 + 2\tau_k(1 - \tau_k)\langle d^k, g^k\rangle + \tau_k^2 |g^k|^2 \\ &\leq (1 - 2(1 - \sigma)\tau_k + \tau_k^2)|d^k|^2 + \tau_k^2 |g^k|^2, \end{aligned}$$

where in the last inequality we used the conditions of Step 2. Proceeding exactly as in the proofs of Lemmas 1.2 and 1.3, we conclude that $d^k \to 0$ and $\{g^k\}$ is bounded. Then the sequence of averages $\{s^k\}$ is bounded, too. Let $\bar{s}$ be any accumulation point of $\{s^k\}$. Since $d^k \to 0$, one must have $\bar{s} \geq 0$. By the continuity of f^*, for the corresponding subsequence we get

$$\begin{aligned} f^*(s^k) \to f^*(\bar{s}) &= \max_x \{\langle \bar{s}, x\rangle - f(x)\} \\ &\geq \max_{x \geq 0} \{\langle \bar{s}, x\rangle - f(x)\} \geq -f_{\min}, \end{aligned}$$

where $f_{\min} = \min_{x \geq 0} f(x)$. Consequently,

$$\liminf f^*(s^k) \geq -f_{\min}. \tag{2.1}$$

This combined with inequality (1.4), in the same manner as in Theorem 1.4, yields the required result. □

We also have an analog of Theorem 1.5.

Theorem 2.2. *Let the assumptions of Theorem 1.4 be satisfied. Then for the sequence of averages*

$$\bar{x}^{k+1} = (1 - \tau_k)\bar{x}^k + \tau_k x^k, \quad k = 0, 1, 2, \ldots,$$

where $\{x^k\}$ *is generated by Algorithm 2, one has*

$$\lim_{k \to \infty} f(\bar{x}^k) = \min_{x \geq 0} f(x).$$

Proof. Proceeding similarly to the proof of Theorem 1.5 we obtain relation (1.5), which implies

$$\limsup \Big(f(\bar{x}^k) + f^*(s^k) \Big) \leq 0. \tag{2.2}$$

On the other hand, $f(\bar{x}^k) \geq f_{\min}$, so we must have $\limsup f^*(s^k) \leq -f_{\min}$. This combined with (2.1) yields

$$\lim_{k \to \infty} f^*(s^k) = -f_{\min}.$$

Our assertion follows now from (2.2). □

3 Applications

Let us discuss some potential applications of the ideas introduced in this paper.

Linear inequalities

Consider the system of linear inequalities

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i, \ i = 1, \ldots, m, \tag{3.1}$$

and the associated optimization problem

$$\min_x \left[f(x) = \max_{1 \leq i \leq m} \left(\sum_{j=1}^{n} a_{ij}x_j - b_i \right) \right].$$

The subproblem solved at Step 1 takes on the form

$$\min_{\mu \geq 0} \max_{1 \leq i \leq m} \left(-\mu \sum_{j=1}^{n} a_{ij}s_j^k - b_i \right).$$

Define the sets

$$J_k^+ = \{j : \sum_{j=1}^{n} a_{ij}s_j^k > 0\},$$

$$J_k^- = \{j : \sum_{j=1}^{n} a_{ij}s_j^k \leq 0\}.$$

If $J_k^- = \emptyset$ then $\sum_{j=1}^{n} a_{ij}s_j^k > 0$ for all i and one can find $\bar{\mu} \geq 0$ such that $-\bar{\mu}s^k$ solves (3.1). It remains to consider the case when $J_k^- \neq \emptyset$ for all k.

If $\mu_k > 0$ there must exist $r \in J_k^-$ and $t \in J_k^+$ such that

$$f(-\mu_k s^k) = -\mu_k \sum_{j=1}^{n} a_{rj}s_j^k - b_r = -\mu_k \sum_{j=1}^{n} a_{tj}s_j^k - b_t.$$

Denote $a_r = (a_{r1}, \ldots, a_{rn})$, $a_t = (a_{t1}, \ldots, a_{tn})$ and define

$$\lambda_k = \frac{\langle a_t, s^k \rangle}{\langle a_t - a_r, s^k \rangle}.$$

Since $a_r \in \partial f(x^k)$, $a_t \in \partial f(x^k)$ and $\lambda_k \in [0, 1]$,

$$g^k = \lambda_k a_r + (1 - \lambda_k) a_t$$

is a subgradient of f at x^k. By the definition of λ_k, $\langle s^k, g^k \rangle = 0$, i.e. g^k satisfies the conditions of Step 2 with $\sigma = 0$.

If $\mu_k = 0$, then there must exist $r \in J_k^-$ such that $b_r \le b_i$, $i = 1, \dots, m$. Taking $g^k = a_r$ we have $\langle g^k, s^k \rangle \le 0$ by the definition of J_k^-.

Constraint aggregation

Consider the convex optimization problem

$$\min h(y) \tag{3.2}$$

$$Ay = b, \tag{3.3}$$

$$y \in Y, \tag{3.4}$$

where $h : \mathbb{R}^m \mapsto \mathbb{R}$ is convex, $Y \subset \mathbb{R}^m$ is convex and compact, A is an $n \times m$ matrix, $b \in \mathbb{R}^n$. Its dual has the form

$$\max f(x), \;\; x \in \mathbb{R}^n,$$

where x is the vector of Lagrange multipliers and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the dual function defined as follows:

$$f(x) = \min_{y \in Y} \{h(y) + \langle x, Ay - b \rangle\}.$$

Clearly, $-f$ is convex and co-finite. Let us apply Algorithm 1 to the dual problem (with obvious modifications reflecting the change from minimization to maximization). Step 1 takes on the form

$$\max_{\mu \ge 0} \min_{y \in Y} \left\{ h(y) + \mu \langle s^k, Ay - b \rangle \right\},$$

which, under appropriate constraint qualification, is equivalent to the following optimization problem

$$\min h(y) \tag{3.5}$$

$$\langle s^k, Ay - b \rangle \le 0, \tag{3.6}$$

$$y \in Y. \tag{3.7}$$

The subgradient g^k satisfying the conditions of Step 2 is given by

$$g^k = Ay^k - b, \tag{3.8}$$

where y^k is the solution of (3.5)-(3.7). Finally, the subgradient averaging rule of Step 3 can be written as

$$z^{k+1} = (1 - \tau_k) z^k + \tau_k y^k, \tag{3.9}$$

$$s^{k+1} = A z^{k+1} - b. \tag{3.10}$$

The algorithm (3.5)-(3.10) can be regarded as an iterative constraint aggregation procedure for solving (3.2)-(3.4): it replaces the constraints (3.3) by a single surrogate inequality (3.6). This idea has been analysed in [2].

If the original problem, instead of (3.3), has inequality constraints

$$Ay \le b,$$

the dual problem has non-negativity constraints on x, so Algorithm 2 applies. The only modification with respect to (3.5)-(3.10) is that (3.10) is replaced by the projection:

$$s^{k+1} = \left(Az^{k+1} - b\right)_+,$$

where $(v_+)_j = \max(0, v_j)$, $j = 1, \ldots, n$. In a similar way we can treat convex inequalities (see [2] for the details missing here, such as the constraint qualification condition, various modifications and extension, analysis of the rate of convergence, etc).

Saddle point seeking

The previous example can be in a straighforward manner generalized to the saddle point problem. Let $L : \mathbb{R}^n \times Y \mapsto R$ be a convex-concave function. Assuming that L is strictly concave in its second argument, we can find a saddle point $(\hat{x}, \hat{y})$ of L in the following way. First, we solve the problem

$$\min_{x \in \mathbb{R}^n} \left[f(x) = \sup_{y \in Y} L(x, y) \right] \tag{3.11}$$

to get $\hat{x}$ and then we define $\hat{y}$ as the maximizer of $L(\hat{x}, \cdot)$ over Y. It turns out that Step 1 of Algorithm 1 applied to (3.11) takes on the form:

$$\min_{\mu \geq 0} \sup_{y \in Y} L(-\mu s^k, y).$$

By defining the function $\Lambda_k(\mu, y) = L(-\mu s^k, y)$ we can equivalently formulate Step 1 as follows: *find a saddle point* (μ_k, y^k) *of* Λ_k *on* $\mathbb{R}_+ \times Y$. Moreover, if L is continuosly differentiable with respect to the first argument, then $g^k = \nabla_x L(-\mu_k s^k, y^k)$ satisfies the conditions of Step 2 with $\sigma = 0$.

References

[1] Yu.M. Ermoliev, *Methods of Stochastic Programming*, Nauka, Moscow, 1976 (in Russian)

[2] Yu.M. Ermoliev, A. Kryazhimskii and A. Ruszczyński, "A constraint aggregation principle in convex optimization", working paper WP-95-015, International Institute for Applied Systems Analysis, Laxenburg, 1995.

[3] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[4] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Springer-Verlag, Berlin, 1985.

[5] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1973).

[6] A. Ruszczyński, "A linearization method for nonsmooth stochastic programming problems", *Mathematics of Operations Research* 12 (1987) 32-49.