Optimization Methods & Software
Vol. 00, No. 0, Month 2009, 1–12

Envelope stepsize control for iterative algorithms based on Fejer processes with attractants

E.A. Nurminski*

Institute for Automation and Control Processes, Far Eastern Branch of RAS,
Far Eastern National University, Vladivostok, Russia

(*Received 24 November 2008; final version received 19 June 2009*)

Fejer processes are frequently used models for many iterative algorithms in optimization and related areas. They can be combined with different kinds of decomposition schemes and generate various projection-type methods suitable for parallel computations. This paper reviews some recent results on Fejer processes with diminishing disturbances and suggests a new adaptive parameter-free stepsize control rule for such algorithms.

Keywords: Fejer processes; convex optimization; decomposition; stepsize control; subgradient method

AMS Subject Classification: 90C25; 90C52; 49M27

1. Introduction

This article mainly describes an attempt to speed up practical convergence of iterative methods for solving optimization and related problems, which can be described as recursive application of Fejer operators. Such algorithmic models are quite general and commonly used in convex feasibility problems (CFP) where Fejer operators can be easily constructed [2]. When certain problem-specific disturbances are added to such processes it is possible to suggest for constrained convex nondifferentiable optimization analogues of CFP-algorithms, ample opportunities for decomposition and parallel computations.

The above mentioned disturbances in the simplest cases are subgradients of the objective function at corresponding points scaled by stepsizes $\lambda_k > 0, k = 0, 1, \dots$ and theoretical convergence is typically guaranteed if $\lambda_m \to +0$ and $\sum_k^m \lambda_k \to \infty$ when $m \to \infty$. However, these assumptions, in practice, result in slow convergence, and it is of practical as well as theoretical interest to speed it up with adaptive rules for stepsizes λ_k. Nondifferentiability of the objective function in convex optimization excludes, however, the use of steepest-descent-like approaches, and there are only a few results on the stepsize control in subgradient algorithms which use the subgradient oracle only, without additional information about objective functions and constraints.

*Email: nurmi@dvo.ru

ISSN 1055-6788 print/ISSN 1029-4937 online

DOI: 10.1080/10556780903151490
http://www.informaworld.com

In this paper, the adaptive stepsize control rule that is based on imitation of the optimality condition is suggested, and its theoretical validation is given. Illustrative numerical experiments demonstrated practically linear convergence of Fejer algorithms with disturbances under this stepsize control.

2. Notations and preliminaries

Let $\mathbb{R}$ be the real axis and E denote a finite dimensional Euclidian space with the inner product xy and the norm $\|x\| = \sqrt{xx}$. The unit ball $\{x{:}\|x\| \leq 1\}$ will be denoted as B. For $x \in E$ and $U \subset E$, the sum $x + U = \{x + u{:}u \in U\}$. The convex hull of a family of vectors $\{a^i : i \in \mathcal{N}\} \subset E$ where $\mathcal{N} = \{1, 2, \dots, N\}$ will be denoted as

$$\mathrm{co}\,\{a^i : i \in \mathcal{N}\} = \left\{ a = \sum_{i=1}^{N} \lambda_i a^i : \sum_{i=1}^{N} \lambda_i = 1, \lambda_i \geq 0, i \in \mathcal{N} \right\} = \{a = A\lambda : \lambda \in \Delta\},$$

where $\Delta = \{\lambda = (\lambda_1, \lambda_2, \dots, \lambda_N) : \sum_{i=1}^{N} \lambda_i = 1, \lambda_i \geq 0, i \in \mathcal{N}\}$—standard simplex, A is the matrix of column vectors a^i, $i \in \mathcal{N}$.

Definition 2.1 *An operator $F{:}E \to E$ is called Fejer (with respect to a given nonempty set V) if for any $v \in V$*

$$\|F(x) - v\| \leq \|x - v\|. \tag{1}$$

It immediately follows from the definition that any $v \in V$ is a fixed point of F. For the relevant F, the corresponding set V will be assumed to be closed throughout the paper.

We define Fejer processes by the following recursive relationship:

$$x^{k+1} = F(x^k), \quad k = 0, 1, \dots \tag{2}$$

where $F(\,\cdot\,)$ is a Fejer operator and x^0 some starting point. The theory of Fejer processes typically answers the question, under which conditions the sequence (2) converges in this or that sense to the set V [10]. To ensure this convergence and for the purpose of this article, the following property that is stronger than Definition 2.1, is necessary.

Definition 2.2 *A Fejer operator F is called locally strong Fejer if for any $\bar{x} \notin V$, there exists a neighbourhood of zero U and $\alpha < 1$ such that $\|F(x) - v\| \leq \alpha\|x - v\|$ for any $v \in V$ and $x \in \bar{x} + U$.*

For the purposes of further applications, we consider a modification of Equation (2):

$$x^{k+1} = F_k(x^k + z^k), \quad k = 0, 1, \dots \tag{3}$$

where z^k is an arbitrary (for a moment) diminishing ($z^k \to 0$) disturbance and F_k is selected from some finite collection $\mathcal{F} = \{\phi_i, i = 1, 2, \dots, M\}$ of locally strong Fejer operators. We are especially interested in the case when such a collection is associated with representation of V as the intersection of $V_i, i = 1, 2, \dots, M$:

$$V = \bigcap_{i=1}^{M} V_i,$$

and each ϕ_i is a locally strong Fejer operator for V_i. It can be shown (see [8] for details) that if for each k operator F_k is one of the ϕ_i, where i is such that $x^k + z^k \notin V_i$, then the presence of diminishing disturbances z^k does not prevent convergence of Equation (3) to V.

By the special choice of disturbances z^s, the sequence $\{x^s\}$ generated by Equation (3) can be forwarded towards a selected subset of V, which will be denoted as $Z \subset V$. For that, we define the notion of restricted attractant as set-valued mapping $\Phi{:}E \to 2^E$, which is directed towards Z at $x \in V$. More precisely,

Definition 2.3 *Set-valued map $\Phi{:}E \to 2^E$ is called a locally restricted attractant of $Z \subset V$ if $g(z - x) \geq 0$ for all $x \in V \setminus Z, g \in \Phi(x)$ and $z \in Z$.*

In fact, as with Fejer operators, we need a stronger definition:

Definition 2.4 *A locally restricted attractant Φ is called a strong locally restricted attractant (of Z), if for each $x' \in V \setminus Z$ there exists a neighbourhood of zero U such that,*

$$g(z - x) \geq \delta > 0,$$

for all $z \in Z, x \in x' + U, g \in \Phi(x)$ and some $\delta > 0$.

When using a special form of disturbances given by attractants, convergence results for Fejer processes can be strengthened.

Theorem 2.5 [8] *Let $\mathcal{F} = \{\phi_i, i = 1, 2, \ldots, M\}$ be a finite family of continuous and locally strong Fejer operators with respect to corresponding V_i and for any $x \notin V = \cap_{i=1}^{M} V_i$ there exists $\iota \in \{1, 2, \ldots, M\}$ such that ϕ_ι is locally strong at x and $D(\cdot)$ is a strong locally restricted attractant of $Z \subset V$. Then the combined process*

$$x^{k+1} = F_k(x^k + \lambda_k d^k),\ d^k \in D(x^k), \quad F_k = \phi_{\iota_k},\ \textit{where } \iota_k \textit{ is such that } x^k + \lambda_k d^k \notin V_{i_k}, \tag{4}$$

if bounded, converges to the set Z if $\lambda_k \to +0$, and $\sum \lambda_k = \infty$.

The immediate application of this approach is to justify the use of sequential or parallel projection in the subgradient projection algorithm

$$x^{k+1} = F_k(x^k - \lambda_k g^k), \quad k = 0, 1, \ldots, \tag{5}$$

when a feasible set X of a convex optimization problem $\min_{x \in X} f(x)$ is possible to represent as an intersection of 'simpler' sets $X = \cap_{i=1}^{M} V_i$. In Equation (5), $F_k(\cdot)$ is the projection operator on a set V_{i_k} from the family of $V_i, i = 1, 2, \ldots, M$ such that $x^k - \lambda_k g^k \not\in V_{i_k}$, and $g^k \in \partial f(x^k)$ is a subgradient of f at x^k. According the Theorem 2.5, there are many possible choices for the selection of V_{i_k}; the round-robin is probably the simplest possible strategy, but picking up the most violated, in some sense, constraint is also possible. It is easy to show that projection on any V_i is a locally strong Fejer operator with respect to V_i, and therefore the Theorem 2.5 guarantees the convergence of such methods (see again [8] for details).

Theorem 2.5 opens many new possibilities for new algorithms of constrained convex optimization, however, the diverging series condition for stepsize λ_k used in the Theorem 2.5 is known to result in slow convergence. Therefore, it is of theoretical as well as practical interest to search for other stepsize control rules with established convergence and better computational performance. In this paper, the adaptive stepsize rule for Equation (4) is suggested and its theoretical convergence is established. Numerical experiments demonstrated quite satisfactory computational performance of this method.

To study the convergence of proposed algorithms, we use convergence conditions that proved to be rather convenient for the analysis of iterative algorithms, especially in the field of nondifferentiable optimization [6]. From the point of view of these conditions, an algorithm is a rule for constructing an infinite sequence of points $\{x^k\}$ that should converge to some target set $X_\star$. This target set may be a solution set of a given optimization or feasibility problem, fixed points of a given operator, set of points satisfying necessary optimality conditions and the like.

Convergence for a certain subsequence of $\{x^k\} \subset E$ is guaranteed if the following conditions are fulfilled:

A1 Sequence $\{x^k\}$ is bounded.
A2 There is a continuous function $W(x) : E \to \mathbb{R}$ such that if $\{x^k\}$ has a limit point $x' \notin X_\star$, then it has another limit point x'' such that $W(x'') < W(x')$.

It is easy to show that under these conditions the sequence $\{x^k\}$ has a limit point $x^\star \in X_\star$. Indeed, denote a set of limit points of the sequence $\{x^k\}$ as $\bar{X}$. It is a closed bounded set and because of continuity of W the set $\bar{W} = \{W(x) : x \in \bar{X}\}$ is closed and bounded as well. Let $w_\star = \{\min w : w \in \bar{W}\}$ and $\{x^{k_t}\}$ be the corresponding subsequence such that $\lim_{t\to\infty} W(x^{k_t}) = w_\star$. Without loss of generality, one can assume that there is a limit $\lim_{t\to\infty} x^{k_t} = x^\star$. Obviously $x^\star \in X_\star$, otherwise according to condition A2 there is another limit point $\bar{x}^\star$ with $W(\bar{x}^\star) < W(x^\star) = w_\star$, which contradicts the definition of $w_\star$.

Conditions A1 and A2 are insufficient, however, to prove that all limit points of $\{x^k\}$ belong to $X_\star$. It is possible to ensure the latter if more stringent monotonicity of the sequence $\{W(x^k)\}$ and specific features of the set $W_\star = \{W(x) \colon x \in X_\star\}$ are requested. The resulting conditions may look like the following:

B1 Sequence $\{x^k\}$ is bounded.
B2 For $\{x^{k_t}\} \to x'$ when $t \to \infty$ with $x' \notin X_\star$, there exists $\epsilon > 0$ such that for any t

$$m_t = \inf\{m : \|x^{k_t} - x^m\| > \epsilon\} < \infty. \tag{6}$$

B3 There is a continuous function $W(x) : E \to \mathbb{R}$ such that

$$\limsup_{t\to\infty} W(x^{m_t}) < \lim_{t\to\infty} W(x^{k_t}) = W(x') \tag{7}$$

for any subsequences $\{x^{k_t}\}, \{x^{m_t}\}$ satisfying B2.
B4 The set $W_\star = \{W(x^\star), x^\star \in X_\star\}$ is such that $\mathbb{R} \setminus W_\star$ is everywhere dense.
B5 If $\{x^{k_t}\} \to x^\star \in X_\star$, then $\|x^{k_t+1} - x^{k_t}\| \to 0$ when $t \to \infty$.

Conditions B2 and B3 imply A2 so it follows from the above that $\{x^k\}$ has at least one limit point in $X_\star$. To prove that there are no limit points out of $X_\star$, it is easy to come to a contradiction.

Indeed, if there is a limit point $x' \notin X_\star$, then according to B2 and B3 there is a second limit point x'' such that $\|x'' - x'\| \geq \epsilon$ and $W(x'') < W(x')$. According to B4, it is possible to select subinterval $[a, b] \subset (W(x''), W(x'))$ such that $a \notin W_\star$. As $\{W(x^k)\}$ infinitely often crosses $[a, b]$, it is possible to select from $\{x^k\}$ subsequences $\{x^{p_t}\}, \{x^{q_t}\}$ such that $p_t < q_t$, $W(x^{p_t}) \leq a$, $W(x^s) \geq a$ for $p_k < s \leq q_t$ and $W(x^{q_t}) \geq (a + b)/2$. Figure 1 may clarify the role of the relevant subsequences.

Without loss of generality, it can be assumed that $x^{p_t} \to \bar{x}'$. It is clear that $\bar{x}' \notin X_\star$ otherwise from $W(x^{p_t}) \leq a < W(x^{p_t+1})$ and $\|x^{p_t+1} - x^{p_t}\| \to 0$ it follows that $W(x^{p_t}) \to a = W(\bar{x}')$ and $\bar{x}'$ cannot belong to $X_\star$ due to the special choice of a.

Define as in Equation (6)

$$r_t = \inf\{r : \|x^{p_t} - x^r\| > \epsilon\} < \infty,$$

where ϵ is sufficiently small that $|W(\bar{x}') - W(x)| \leq (b - a)/4$ for all $\|\bar{x}' - x\| < 4\epsilon$. For t large enough $\|x^{p_t} - \bar{x}'\| \leq \epsilon$ and

$$\|x^s - \bar{x}'\| = \|x^{p_t} - \bar{x}' - x^{p_t} + x^s\| \leq \|x^{p_t} - \bar{x}'\| + \|x^{p_t} - x^s\| \leq 2\epsilon,$$

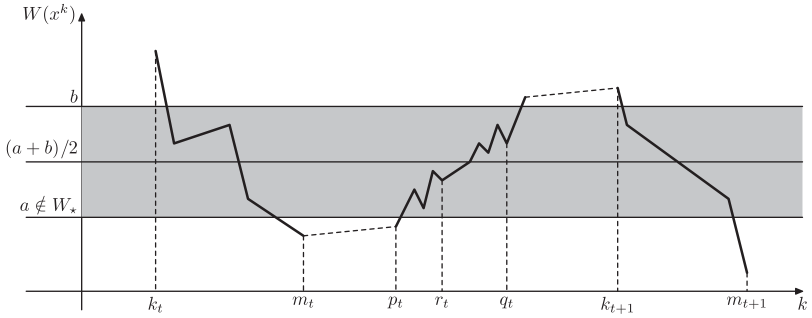

Figure 1. Subsequences involved: $x^{k_t} \to x' \notin X_\star$, $\|x^{m_t} - x^{k_t}\| > \epsilon$, $\limsup_{t\to\infty} W(x^{m_t}) \le a < b \le \lim_{t\to\infty} W(x^{k_t})$, $x^{p_t} \to \bar{x}' \notin X_\star$, $\|x^{p_t} - x^{r_t}\| > \epsilon$,$\limsup_{t\to\infty} W(x^{r_t}) \ge \lim_{t\to\infty} W(x^{p_t}) = W(\bar{x}')$.

for all $p_t < s < r_t$ and consequently $|W(x^s) - W(\bar{x}')| \le (b-a)/4$. Hence $W(x^s) \le W(\bar{x}') + (b-a)/4 < a + (b-a)/2 = (a+b)/2 \le W(x^{q_t})$ for s such that $p_t \le s < r_t$, and therefore $r_t \le q_t$. By construction, $W(x^s) \ge a$ for $p_t < s \le q_t$ and hence $W(x^{r_t}) \ge a$ and

$$\limsup_{t\to\infty} W(x^{r_t}) \ge a \ge W(\bar{x}'),$$

which contradicts B3. This contradiction proves that all limit points of the sequence $\{x^k\}$ belong to $X_\star$.

The ideas of this approach can be traced down to Lyapunov conditions for continuous-time dynamical systems with Equation (7) being the analogue of the negative sign of full derivative of a Lyapunov function along the trajectory of system, described by ordinary differential equations. For this reason, we will sometimes call $W(\cdot)$ a Lyapunov function of the process $\{x^k\}$.

The advantage of using conditions B1–B5 consists of separating local analysis of the limit behaviour of an algorithm in a vicinity of 'non-optimal' point (B2, B3) from checking out global conditions (B1, B4, B5). Global conditions depend essentially on properties of a Lyapunov function W and even as no definite recipes for constructing such functions exist, the objective function itself and the distance to the optimum are typical choices. Weak monotonicity required by B3 makes it easier to prove even in the cases when algorithms are nonmonotone, as we see in the examples.

The development of conditions A1 and A2 and B1–B5 were inspired by the pioneering work of Zangwill [11], but they were especially tailored to deal with nonmonotone algorithms of nondifferentiable optimization. Among related work the 'gradient related' algorithms of Bertsekas [5] also bear some relation to B3 and B4 with $W(x) = f(x)$, but as the name suggests, it requires differentiability of the objective function.

3. Envelope stepsize control

As numerical experiments and theoretical analysis show, the diverging sum series stepsize rule used in the Theorem 2.5 and in many theoretical studies of subgradient-like algorithms of nondifferentiable optimization as well results in slow convergence. Here, we present a simple and rather general idea for stepsize control in methods, based on Fejer processes with attractants. For simplicity, we consider the iterative process

$$x^{k+1} = x^k - \lambda_k d^k, \quad d^k \in D(x^k), \tag{8}$$

where $D(x)$ is a set-valued attractant whose properties will be specified later. To simplify notations let $D(p, q) = \mathrm{co}\,\{d^p, d^{p+1}, \dots, d^q\}$.

For a given sequence $\theta_m \to +0, m = 0, 1, \dots$ determine corresponding sequences of indices $\{k_m\}$ and numbers $\{\lambda_k\}$ by the following recursive relationships:

- Set $k_0 = 0$ and pick up initial $\lambda_0 > 0$. Let $q \in (0, 1)$.
- For given m and k_m, determine k_{m+1} as the index that satisfies conditions

$$0 \notin D(k_m, k) + \theta_m B, \quad k_m \le k < k_{m+1}, \quad 0 \in D(k_m, k_{m+1}) + \theta_m B \tag{9}$$

with $\lambda_k = \lambda_{k_m}$ for $k_m \le k < k_{m+1}$. Set

$$\lambda_{k_{m+1}} = q\lambda_{k_m}. \tag{10}$$

In other words, condition (9) detects the first instance when $\{x^k\}$ seems to start cycling, λ_k is kept constant between k_m and k_{m+1} and, according to Equation (10), at $k = k_{m+1}$ stepsize is diminished by factor q. This idea, that is to keep stepsize constant while we seem to be moving in a certain direction and decrease it otherwise, is by no means new and can be traced back as far as Armijo [1]. Recently, as a certain heuristic to improve a current approximate solution, it was propagated in [3] with successful applications in image processing in tomography.

3.1 *Convergence*

The following theorem establishes convergence of the process (8) with the stepsize rules (9) and (10). Denote $X_\star = \{x^\star : 0 \in D(x^\star)\}$. The following theorem holds.

Theorem 3.1 *Let D(x) be a convex-valued, locally bounded upper-semi-continuous set-valued locally strong attractant of $X_\star$. Then if the sequence $\{x^k\}$ generated by Equation* (8) *is bounded then all its limit points belong to $X_\star$.*

Proof Note that ks such that $d^k = 0$ can be deleted from Equation (8) and $d^k = 0$ for all k represent the trivial case. So assume that $d^k \neq 0$ for all k. Next, we establish that the sequence $\{k_m\}$, defined by Equation (9) is infinite, or, in other words, λ_k is decreased in accordance with Equation (10) infinitely many times.

Indeed, if λ_k is decreased only a finite number of times, then there is M such that for all $k > k_M$

$$0 \notin D(k_M + 1, k) + 2\delta_M B = D_k + 2\delta_M B$$

for some $\delta_M > 0$. By monotonicity, D_k with respect to inclusion there is a Kuratowski limit (see [4] for the definition) $\lim_{k\to\infty} D_k = \tilde{D}$ with $0 \notin \bar{G} + \delta_M B$, where $\bar{D}$ is the closure of $\tilde{D}$. Sets $\tilde{D}$ and $\bar{D}$ are of course convex, and therefore there exists $v \in \bar{D}$ such that

$$v\bar{d} \ge \|v\|^2,$$

for all $\bar{d} \in \bar{D} + \delta_M B$. By representing $\bar{d}$ as $d + \delta_M z, d \in \bar{D}, z \in B$, obtain

$$vd \ge \|v\|^2 - \delta_M vz,$$

for all $z \in U$. After taking supremum of the right-hand side with respect to $z \in B$ obtain

$$vd \ge \|v\|^2 + \delta_M \|v\| > \delta_M \|v\| > 0$$

or

$$\bar{v}d > \delta_M, \tag{11}$$

where $\bar{v} = v/\|v\|$. As $\bar{D} \supset D_k$ for all $k \geq k_M$ the inequality (11) holds for all $d^k, k > k_M$, and hence

$$\|x^{k_M} - x^K\| \geq (x^{k_M} - x^K)\bar{v} = \sum_{k=k_M}^{K-1} \lambda_M d^k \bar{v} \geq \lambda_M(K - 1 - k_M)\delta_M \longrightarrow \infty$$

when $K \to \infty$, which contradicts the boundness of $\{x^k\}$. It proves that $\lambda_k \to 0$ and also $\|x^{k+1} - x^k\| \to 0$ and hence B4 is fulfilled.

In what follows, we show that B2 is fulfilled as well. Assume that $\{x^{n_k}\}$ is a certain subsequence, which converges to $x' \notin X_\star$. Then $0 \notin D(x')$, and by upper-semicontinuity of $D(\cdot)$ there exists $\epsilon, \delta > 0$ such that

$$0 \notin \text{co}\,\{D(x) : \|x' - x\| \leq 4\epsilon\} + \delta B. \tag{12}$$

Consider n_k large enough that $\|x^{n_m} - x'\| \leq \epsilon$ for $m \geq k$. Without loss of generality, we can assume that the corresponding $2\theta_m < \delta$. Then if $\{x^l\}$ remains in the 4ϵ-neighbourhood of x' for $l > n_k$, then not more than one change in the value of λ_k can occur. In other words, among indices l such that $l \geq n_k$ and $\|x^l - x'\| \leq 4\epsilon$ there is not more than one $l \in \{k_m\}$. In fact, if there were two such indices $k_{m'}$ and $k_{m'+1}$ it would contradict Equation (12):

$$0 \in D(k_{m'}, k_{m'+1}) \subset \text{co}\,\{D(x) : \|x' - x\| \leq 4\epsilon\} + \delta B.$$

Therefore the assumption that $\{x^l : l \geq n_k\} \in \{x : \|x - x'\| \leq 4\epsilon\}$ contradicts the infiniteness of $\{k_m\}$ and hence for all k there exist $m_k \geq n_k$ such that

$$\|x^{m_k} - x^{n_k}\| > \epsilon, \quad \|x^l - x^{n_k}\| \leq \epsilon \text{ for all } l \text{ such that } n_k \leq l < m_k.$$

Finally, we show that B3 is fulfilled as well. Assume as above that the sequence $\{x^k\}$ has a subsequence $\{x^{n_k}\} \to x' \notin X_\star$. Then, according to B2 for any $\epsilon > 0$ small enough there exists $\{x^{m_k}\}$ such that for all k

$$\|x^{n_k} - x^s\| \leq \epsilon, \quad n_k < s \leq m_k, \quad \|x^{m_k} - x^{n_k}\| > \epsilon.$$

Estimate $W(x^{m_k}) = \min_{x^\star \in X_\star} \|x^{m_k} - x^\star\|^2$ from above as follows:

$$\begin{aligned} W(x^{m_k}) &\leq \|x^{m_k} - x^\star\|^2 = \|x^{m_k} - x^{n_k} + x^{n_k} - x^\star\|^2 \\ &= \|x^{n_k} - x^\star\|^2 + 2(x^{n_k} - x^\star)(x^{m_k} - x^{n_k}) + \|x^{m_k} - x^{n_k}\|^2 \\ &\leq \|x^{n_k} - x^\star\|^2 + 2(x^{n_k} - x^\star)(x^{m_k} - x^{n_k}) + \epsilon^2 \end{aligned}$$

for any $x^\star \in X_\star$. Taking into account that

$$x^{m_k} - x^{n_k} = \sum_{s=n_k}^{m_k-1} (x^{s+1} - x^s) = \sum_{s=n_k}^{m_k-1} \lambda_s d^s,$$

obtain

$$
\begin{aligned}
W(x^{m_k}) &\le \|x^{n_k} - x^\star\|^2 + 2\sum_{s=n_k}^{m_k-1} \lambda_s (x^{n_k} - x^\star) d^s + \epsilon^2 = \|x^{n_k} - x^\star\|^2 \\
&\quad + 2\sum_{s=n_k}^{m_k-1} \lambda_s (x^{n_k} - x^s + x^s - x^\star) d^s + \epsilon^2 \le \|x^{n_k} - x^\star\|^2 + 2\sum_{s=n_k}^{m_k-1} \lambda_s (x^s - x^\star) d^s \\
&\quad + 2\sum_{s=n_k}^{m_k-1} \lambda_s \|x^{n_k} - x^s\| \|d^s\| + \epsilon^2 \le \|x^{n_k} - x^\star\|^2 + 2\sum_{s=n_k}^{m_k-1} (x^s - x^\star)\lambda_s d^s \\
&\quad + 2\epsilon C \sum_{s=n_k}^{m_k-1} \lambda_s + \epsilon^2,
\end{aligned}
$$

where C is some constant large enough. As $D(\,\cdot\,)$ is an attractant $(x^s - x^\star)d^s \le -\gamma$ for some $\gamma > 0$ and hence

$$
W(x^{m_k}) \le \|x^{n_k} - x^\star\|^2 - 2\gamma \sum_{s=n_k}^{m_k-1} \lambda_s + 2\epsilon C \sum_{s=n_k}^{m_k-1} \lambda_s + \epsilon^2.
$$

Assuming $\epsilon < 2\gamma/C,$ the last inequality can be strengthened to

$$
W(x^{m_k}) \le \|x^{n_k} - x^\star\|^2 - \gamma \sum_{s=n_k}^{m_k-1} \lambda_s + \epsilon^2. \tag{13}
$$

The sum $\sum_{s=n_k}^{m_k-1} \lambda_s$ can be estimated from below

$$
\epsilon < \|x^{m_k} - x^{n_k}\| \le \sum_{s=n_k}^{m_k-1} \lambda_s \|d^s\| \le \frac{1}{2} C \sum_{s=n_k}^{m_k-1} \lambda_s
$$

and after substituting that in Equation (13) obtain

$$
W(x^{m_k}) \le \|x^{n_k} - x^\star\|^2 - \frac{2\gamma\epsilon}{C} + \epsilon^2 \le \|x^{n_k} - x^\star\|^2 - \frac{\gamma\epsilon}{C}
$$

for arbitrary $x^\star \in V$. Computing the infinum of the right-hand side with respect to $x^\star \in X_\star$ yields:

$$
W(x^{m_k}) \le W(x^{n_k}) - \frac{\gamma\epsilon}{C}.
$$

Passing to the limit when $k \to \infty$ results in

$$
\limsup_{k\to\infty} W(x^{m_k}) \le \lim_{k\to\infty} W(x^{n_k}) - \frac{\gamma\epsilon}{C} = W(x') - \frac{\gamma\epsilon}{C} < W(x'),
$$

which proves B3 and hence completes the proof of the theorem. ■

A few words about computational issues related to the practical use of envelope stepsize control (ESC). To apply this stepsize rule we need to perform repetitive checks of inclusion

$$
0 \in D(k_m, k) + \delta_m B, \quad k = k_m + 1, k_m + 2, \dots. \tag{14}
$$

A natural way to check Equation (14) is to solve the least norm problem for the set $D(k_m, k)$ and compare its result with δ_m. This is a nontrivial problem, however, the algorithm [7] demonstrated in our tests quite adequate performance. Much can also be gained from the incremental growth of the set $D(k_m, k)$, when the least norm solution on the previous iteration can be used as a good starting point for the next, which can be easily incorporated in the algorithm [7].

3.2 *Illustrative examples*

It is useful to consider even tiny illustrative examples of the use of ESC, and here we consider two of them: unconditional convex optimization and nonlinearly constrained. Despite their different natures ESC demonstrated in both cases similar computational behaviour and outperformed the previously suggested [9].

3.2.1 *Unconstrained convex optimization*

Let us apply the subgradient method for minimization of the piece-wise linear function

$$f(x) = \max\{y_i, i = 1, 2, 3\}, \quad (y_1, y_2, y_3) = y = Ax \tag{15}$$

of 2-dimensional vector $x = (x_1, x_2)$ defined by the matrix

$$A = \begin{vmatrix} 1 & 0 \\ -0.5 & -0.4 \\ -3 & 0.2 \end{vmatrix}.$$

The origin $x^\star = (0, 0)$ with $f(x^\star) = 0$ is a trivial solution of this problem. The algorithm

$$x^{k+1} = x^k - \lambda_k g^k g^k \in \partial f(x^k), \quad k = 0, 1, \dots \tag{16}$$

was started from the initial point $x^0 = (5, 7)$ with rather large initial stepsize $\lambda_0 = \|x^0\| = 8.6$. The subgradient algorithm (16) can be considered as a very special case of a Fejer process with diminishing disturbances (4) when F_k is just an identity operator, $V = E, Z = \{x^\star : 0 \in \partial f(x^\star)\}$ and the attractant mapping $D(\cdot)$ is the subdifferential of the objective function (15).

The results of application of ESC are shown in Figure 2. We can clearly observe the linear rate of convergence with some perturbations, and it is clear also that the algorithm is not monotone either in terms of the objective function or in terms of the Euclidian distance to the optimal point, which is a traditional performance indicator of convex optimization. Therefore the proof of convergence for such algorithms can be obtained only by nontraditional argument-like conditions B1–B5 and also the very notion of the rate of convergence should be modified to cover such cases.

It may be worth noting that the ESC stepsize rule does not make any assumptions about the structure of the objective function and is strictly 'subgradient oracle-based'. There are not so many stepsize recommendations for such a case and it is interesting to compare these computational

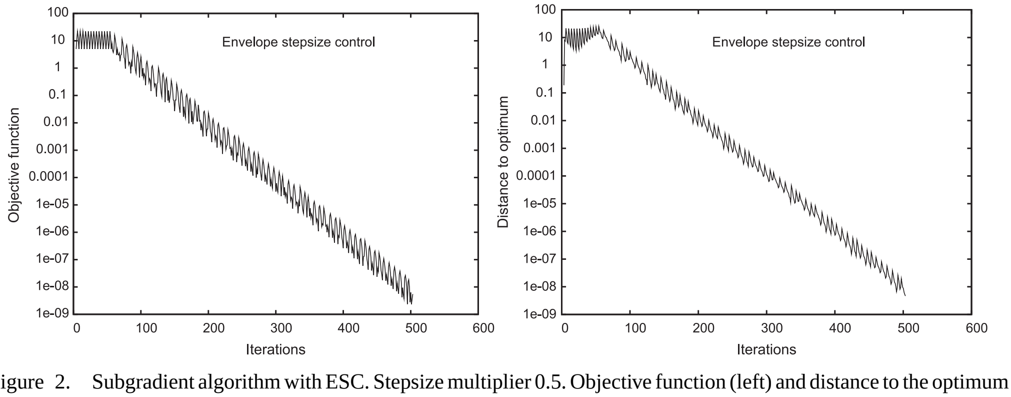

Figure 2. Subgradient algorithm with ESC. Stepsize multiplier 0.5. Objective function (left) and distance to the optimum (right).

results with theoretical estimates [9] for another stepsize rule that guarantees a linear rate of convergence. In this work, the subgradient method with normalized subgradient

$$x^{k+1} = x^k - \frac{h_k g^k}{\|g^k\|}, \quad k = 0, 1, \ldots, \quad g^k \in \partial f(x^k)$$

is considered and the use of $h_{k+1} = \sin(\phi) h_k = q_f h_k$, $h_0 \geq \|x^0 - x^\star\| \mathrm{co}\,s(\phi)$ is proposed. The angle ϕ is determined from the condition that for any $x \neq x^\star$ and $g \in \partial f(x)$ the following inequality holds

$$g(x - x^\star) \geq \cos(\phi) \|g\| \|x - x^\star\| \tag{17}$$

with $\phi \in [\pi/4, \pi/2)$. For the function (15), the maximal angle ϕ is determined by the subdifferential of this function on the line $y_1 = y_3$, which is equal to the convex hull of the first and the third rows of A. By direct computation for Equation (17), obtain $\cos(\phi) = 0.016609$, which gives $q_f = 0.99986$. After the same 500 iterations the initial stepsize will be decreased to only $0.93335 h_0$, which in fact means that no practical convergence occurs.

3.2.2 *Quadratically constrained convex optimization*

Another example is given by the convex optimization problem with quadratic constraints

$$\begin{aligned} \min f(x) &= \min \max_{i=1,2,3} \sum_{j=1}^{3} a_{ij} x_j \\ \|x - \mathrm{e}^1\|^2 - 4 &= h_1(x) \leq 0, \\ \|x - \mathrm{e}^2\|^2 - 4 &= h_2(x) \leq 0, \end{aligned} \tag{18}$$

where vectors $\mathrm{e}^1 = (1, 0, 0)$, $\mathrm{e}^2 = (-1, 0, 0)$, and the matrix $A = \|a_{ij}\|$ has the following entries:

$$A = \left| \begin{array}{rrr} -1 & -1 & 2 \\ 2 & 1 & 3 \\ 1 & 4 & 1 \end{array} \right| .$$

The feasible set X in this problem is the intersection of two balls

$$X = B_1 \cap B_2, \quad B_1 = \{x : h_1(x) \leq 0\}, \quad B_2 = \{x : h_2(x) \leq 0\}$$

and the projection on any of them presents no difficulty. To make use of this we apply for solution of Equation (18) the Fejer process

$$x^{k+1} = F_k(x^k - \lambda_k g^k), \quad k = 0, 1, \ldots, \quad x^0 = (1, 2, 3) \tag{19}$$

with the attractant $g^k \in \partial f(x^k)$ is the subgradient of f, computed at x^k. The Fejer operator F_k is constructed with the help of projection operators Π_i defined as

$$\|\Pi_i(z) - z\| = \min_{x \in B_i} \|x - z\|, \quad i = 1, 2;$$

and

$$F_k(x^k - \lambda_k g^k) = \Pi_{\iota_k}(x^k - \lambda_k g^k),$$

where ι_k is such that

$$h_{\iota_k}(x^k - \lambda_k g^k) = \max\{h_1(x^k - \lambda_k g^k), h_2(x^k - \lambda_k g^k)\} > 0.$$

If $\max\{h_1(x^k - \lambda_k g^k), h_2(x^k - \lambda_k g^k)\} \leq 0$ then F_k is identity.

In other words, the shifted point $x^k - \lambda_k g^k$ is projected on the most violated constraint. In fact it does not matter too much, projection on any violated constraint will result in convergence of the algorithm, possibly slower.

To cast Equation (19) into the form of Equation (8) one has to set $d^k = (F_k(x_k - \lambda_k g^k) - x^k)/\lambda_k = (\bar{x}^k - x^k)/\lambda_k$. The boundness of $\{x^k\}$ already guarantees that $\lambda_k \to 0$ which, according to Theorem 3.1, establishes convergence of $\{x^k\}$ to the feasible set X. It can be shown furthermore that for X with nonempty interior, which is the case, $d^k \in (\mathrm{co}(\partial f(x^k)) + F_k^+) \cap \gamma B = D(x^k)$ where F_k^+ is a positive cone for the cone of feasible directions at x^k and γ is large enough. Then $0 \in D(x^\star), x^\star \in X$ corresponds to the optimality conditions of problem (18).

In Figure 3, convergence of objective function values towards the optimal value -3.05714794 and solutions themselves towards optimal point $x^\star = (0, -0.339683151044523, -1.698415543040009)$ are shown. Again as for the unconstrained optimization problem considered earlier, we observe the linear rate of convergence on the average with marked nonmonotonicity both in terms of objective function and distance to the optimum.

It is interesting to see the dynamics of the stepsize for this problem shown in Figure 4. It can be seen that the stepsize decreases in a more or less regular way, which accounts for the linear

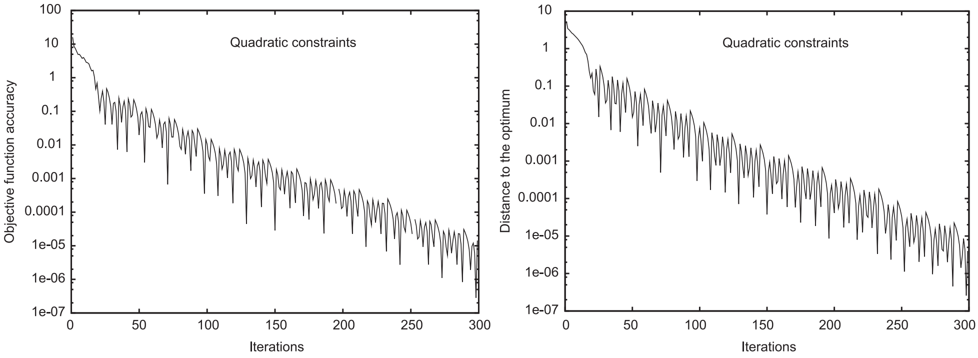

Figure 3. Convergence of the objective function and the distance to the optimum.

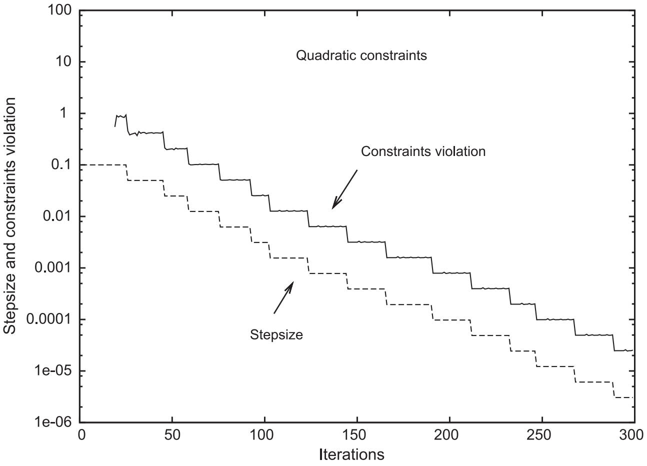

Figure 4. Constraints violation and stepsize.

rate of convergence. Curiously enough, the dynamics of maximal constraint violation also shown in this figure closely parallels the stepsize, however, small deviations from monotonicity can be observed here as well.

4. Conclusions

In general, it can be concluded that a combination of Fejer processes and problem-specific attracting mappings can be used to suggest new algorithms with new opportunities for decomposition and parallel computations. For these algorithms it is possible to use parameter-free adaptive envelope stepsize control for which numerical experiments demonstrated a nonmonotone but close-to-linear rate of convergence.

Acknowledgements

The author wishes to thank G.T. Herman and two anonymous referees for a great amount of meticulous reading of preliminary versions of this paper and many constructive suggestions.

The study was supported by RFBR grant 09-01-90413-Ukr_f_a.

References

[1] L. Armijo, *Minimization of functions having Lipschitz-continuous first partial derivatives*, Pacific J. Math. 16 (1966), pp. 1–3.

[2] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev. 38(3) (1996), pp. 367–426.

[3] D. Butnariu, R. Davidi, G.T. Herman, and I.G. Kazantsev, *Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems*, IEEE J. Sel. Topics Signal Process 1 (2007), pp. 540–546.

[4] K. Kuratowski, *Topology* Vol. I. New edition, revised and augmented, Translated from the French by J. Jaworowski Academic Press, New York-London; Państwowe Wydawnictwo Naukowe, Warsaw, 1966.

[5] A. Nedić and D.P. Bertsekas, *Incremental subgradient methods for non-differentiable optimization*, SIAM J. Optim. 12 (2001), pp. 109–138.

[6] E.A. Nurminski, *Numerical methods of convex optimization*, Nauka, Moscow, 1991, [in Russian].

[7] E.A. Nurminski, *Convergence of the suitable affine subspace method for finding the least distance to a simplex*, Comput. Math. Math. Phys. 45(11) (2005), pp. 1915–1922.

[8] E.A. Nurminski, *The use of additional small disturbances in Fejer models of iterative algorithms*, Zhurn. Vychisl. Mathem. Matem. Physiki [In Russian], 48(12) 2008, pp. 2121–2128.

[9] N.Z. Shor and P.R. Gamburd, *On convergence of generalized gradient descent method*, Kibernetika 6 [in Russian], (1971), pp. 82–84.

[10] V.V. Vasin and I.I. Eremin, *Operators and Fejer Iterative Processes: Theory and Applications*, Ural. Otd. Ross. Acad. Nauk, Yekaterinburg [in Russian], 2005.

[11] W.I. Zangwill, *Convergence conditions for nonlinear programming algorithms*, Manag. Sci. 16(1) (1969), pp. 1–13.