JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 98, No. 3, pp. 663-680, SEPTEMBER 1998

# Error Stability Properties of Generalized Gradient-Type Algorithms[1]

M. V. SOLODOV[2] AND S. K. ZAVRIEV[3]

Communicated by Z. Q. Luo

**Abstract.** We present a unified framework for convergence analysis of generalized subgradient-type algorithms in the presence of perturbations. A principal novel feature of our analysis is that perturbations need not tend to zero in the limit. It is established that the iterates of the algorithms are attracted, in a certain sense, to an $\epsilon$-stationary set of the problem, where $\epsilon$ depends on the magnitude of perturbations. Characterization of the attraction sets is given in the general (nonsmooth and nonconvex) case. The results are further strengthened for convex, weakly sharp, and strongly convex problems. Our analysis extends and unifies previously known results on convergence and stability properties of gradient and subgradient methods, including their incremental, parallel, and heavy ball modifications.

**Key Words.** Error stability, perturbation analysis, gradient-type methods, incremental algorithms, approximate solutions.

## 1. Introduction

We consider the general optimization problem

$$\min_{x \in X} f(x), \tag{1}$$

where $X$ is a convex compact set in $\Re^n$. For the objective function, we assume that $f$: $(X + \tau\mathbb{B}) \to \Re$ is at least Lipschitz continuous and regular (in

[1]The first author was supported by CNPq Grant 300734/95-6. Research of the second author was supported in part by International Science Foundation Grant NBY000, International Science Foundation and Russian Government Grant NBY300, and Russian Foundation for Fundamental Research Grant N95-01-01448.

[2]Assistant Professor, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil.

[3]Professor, Operations Research Department, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia.



0022-3239/98/0900-0663$15.00/0 

the sense of Clarke, Ref. 1) on $X+\tau\mathbb{B}$ for some $\tau\in(0, +\infty]$, where $\mathbb{B}$ is the closed unit ball in $\Re^n$.

Let $X_{\text{opt}}$ and $X_s$ denote the optimal and stationary sets of problem (1) respectively:

$$X_{\text{opt}} := \left\{x\in X \,\middle|\, f(x)=\min_{y\in X} f(y)\right\},$$

$$X_s := \{x\in X \mid 0\in\partial f(x)+N_X(x)\},$$

where $\partial f(x)$ is the set of all generalized gradients (in the sense of Clarke, Ref. 1) of $f(\cdot)$ at $x$ and $N_X(x)$ is the normal cone to the set $X$ at the point $x\in X$. We define the $\epsilon$-stationary set of problem (1) as follows:

$$X_s(\epsilon) := \{x\in X \mid 0\in\partial f(x)+N_X(x)+\epsilon(x)\mathbb{B}\},$$

where $\epsilon: X\to\Re_+$ is a nonnegative upper-semicontinuous function. Clearly, $X_s=X_s(0)$. The $\epsilon$-optimal set of (1) is defined by

$$X_{\text{opt}}(\epsilon) := \left\{x\in X \,\middle|\, f(x)\le \min_{y\in X} f(y)+\epsilon(x)\right\},$$

where again $\epsilon: X\to\Re_+$ some nonnegative upper-semicontinuous function. Obviously, $X_{\text{opt}}(0)=X_{\text{opt}}$.

The primary objective of this paper is to lay down a theoretical framework for convergence analysis of the class of generalized subgradient-type methods in the presence of bounded perturbations. Starting with a current iterate $x^i$, the generalized subgradient projection method computes the next iterate $x^{i+1}$ according to the following formula:

$$x^{i+1} := P_X[x^i-\eta_i(g_i+\delta(x^i))], \qquad g_i\in\partial f(x^i), \qquad \eta_i>0,$$

where $\delta(x^i)$ represents the perturbations (noise) at $x^i$ and $P_X[\cdot]$ denotes the orthogonal projection map onto $X$. In the case of convex $f(\cdot)$, this method is essentially equivalent to the $\epsilon$-subgradient algorithm, because the $\epsilon$-subgradient contains, and is contained in, the image through $\partial f$ of appropriate balls around the given point. In the analysis of such methods, it is typically assumed that the $\epsilon$-subgradients become asymptotically exact (Ref. 2). In this paper, we do not assume convexity and assume that the error terms are merely bounded.

It is clear that, in the presence of bounded perturbations, the iterates generated by the methods in consideration need not converge to the exact stationary set of the problem. In this paper, we show that the iterates are attracted, in a certain sense, to an $\epsilon$-stationary set introduced above (see Theorem 3.1). We give a precise characterization of $\epsilon(\cdot)$ in terms of the

asymptotic behavior of perturbations. Our analysis is based on the novel technique presented in Ref. 3. This approach allows us to deal with essentially perturbed problems [i.e., problems with nonvanishing noise: $\delta(x^i) \not\to 0$ as $i \to \infty$], as well as analyze algorithms that are inherently nonmonotone (e.g., the incremental methods described below). Analysis for perturbed optimization algorithms in a different context (for differentiable functions and different stepsize rules) can be found in Refs. 4–6.

The incremental algorithms (Ref. 7, Section 1.5.2) considered in this paper are designed for minimizing an additive objective function,

$$\min_{x \in X} f(x, \alpha_0) := \sum_{j=1}^{K} f_j(x, \alpha_0), \tag{2}$$

in the case where $K$ is large. These methods work by processing partial objective functions $f_j(\cdot, \alpha_0)$, $j = 1, \ldots, K$, one at a time and immediately updating the variables. Starting with $x^i$, the next iterate $x^{i+1}$ is obtained by the following procedure:

$$z^0 = x^i, \qquad z^1 = z^0 - \eta_i(g_1 + \delta_1), \ldots,$$
$$z^K = z^{K-1} - \eta_i(g_K + \delta_K), \qquad x^{i+1} = P_X[z^K],$$

where $g_j \in \partial f_j(z^{j-1}, \alpha)$, $j = 1, \ldots, K$, and $\delta_j$ are corresponding error terms. The functions $f_j : (X + \tau\mathbb{B}) \times A \to \Re$ depend on a parameter $\alpha \in A \subset \Re$ that may vary during the optimization process. We assume that the set $A$ is bounded and the functions $f_j(\cdot, \alpha)$ are Lipschitz continuous and regular on an open neighborhood of $X + \tau\mathbb{B}$ for every $\alpha \in A$. Problems of the form (2) arise, for example, in least-norm minimization. Of particular importance are machine learning (Ref. 8) and control (Ref. 9) applications, where the computational significance of incremental algorithms is well documented. Among some important applications that involve parameters in the objective function, we note adaptive smoothing techniques (Ref. 10) and neural network training (Refs. 11, 12). It should be noted that the incremental algorithms are inherently nonmonotone (even in the smooth and noise-free case), which makes standard Lyapunov-type convergence analysis techniques inapplicable. Thus, until recently, there has been no rigorous mathematical analysis for incremental algorithms. First deterministic convergence results for noise-free smooth incremental gradient methods are given in Refs. 13–14; the extended Kalman filter (an incremental least-square algorithm) is analyzed in Ref. 15. Other recent work on incremental gradient methods includes Refs. 16–18. In this paper, we fill some of the remaining theoretical gaps, specifically in the perturbation analysis of these methods (Theorems 3.1, 3.2). As a byproduct, we obtain also some new results for the basic generalized

subgradient projection method in the presence of bounded perturbations (see Theorem 4.1), thus improving on Refs. 19–21.

We now describe the notation and some concepts employed in the paper. We define the optimality function $r: X \to \Re_+$ by

$$r(x) := \{\min \|h\| \mid h \in \partial f(x) + N_X(x)\}. \tag{3}$$

It is clear that $r(\cdot)$ is an optimality function for problem (1) in the sense that

$$r(x) = 0, \qquad \text{if } x \in X_s,$$

$$r(x) > 0, \qquad \text{otherwise.}$$

From the definitions of $X_s(\epsilon)$ and $r(x)$, we immediately obtain the following key relation:

$$X_s(\epsilon) = \{x \in X \mid r(x) \leq \epsilon(x)\}. \tag{4}$$

Let $\mathcal{F}(\cdot, \cdot): \mathcal{N} \times X \to \mathcal{M}(\Re^m)$ be a point-to-set mapping (or a multifunction), where $\mathcal{M}(C)$ denotes the set of all subsets of a set $C$ and $\mathcal{N}$ denotes the nonnegative integers. The upper topological limit of $\mathcal{F}(\cdot, \cdot)$ at $x \in \Re^n$ is defined by

$$\operatorname*{\overline{lt}}_{\substack{x'(\in X) \to x \\ i \to \infty}} \mathcal{F}(i, x') := \left\{ y \in \Re^m \;\middle|\; \begin{array}{l} \text{there exist sequences } \{x_i'\}, \{m_i\}, \{y_i\} \\ \text{such that } y_i \in \mathcal{F}(m_i, x_i'),\ i = 1, 2, \ldots, \\ \{x_i'\} \to x,\ \{m_i\} \to \infty, \text{ as } i \to \infty, \\ \text{and } y = \lim_{i \to \infty} y_i \end{array} \right\}.$$

In particular, for a bounded sequence $\{x_i\} \subset X$, $\overline{\text{lt}}_{i \to \infty}\{x_i\}$ is the set of all accumulation points of $\{x_i\}$. We say that a sequence $\{x_i\}$ converges into a set $C$ if $\overline{\text{lt}}_{i \to \infty}\{x_i\} \subset C$.

Note that, under our assumptions, for all $x \in X$ we have

$$\operatorname*{\overline{lt}}_{x'(\in X) \to x} N_X(x') = N_X(x), \tag{5a}$$

$$\operatorname*{\overline{lt}}_{\substack{x'(\in X) \to x \\ \alpha(\in A) \to \alpha_0}} \partial f_j(x', \alpha) \subset \partial f_j(x, \alpha_0). \tag{5b}$$

The rest of the paper is organized as follows. In Section 2, we outline the generalized Lyapunov direct method for stability analysis. In Section 3, we establish the convergence properties of the generalized subgradient projection method and its modifications in the presence of data perturbations. Section 4 contains some new results for a number of special cases. These include the case of asymptotically (relatively) small perturbations and convex, weakly sharp, and strongly convex problems.

## 2. Generalized Lyapunov Direct Method

In this section, we outline the novel convergence analysis technique that was first proposed in Ref. 3 (albeit in a slightly different form). This very useful technique can be viewed as a generalization of the Lyapunov direct method for convergence analysis of nonlinear iterative processes. The classical Lyapunov direct method is a powerful tool for stability analysis of both continuous-time and discrete-time processes (Refs. 22–24). Roughly speaking, this approach reduces the analysis of stability properties of a process to the analysis of local improvement of this process with respect to some scalar criterion $V(\cdot)$, usually called the Lyapunov function. In the classical approach, $V(\cdot)$ decreases monotonically from one iterate of the process to the next. Some typical choices for $V(\cdot)$ are the objective function being minimized or the norm of some optimality measure (see Ref. 24). It should be noted that, in certain situations, for example in the presence of perturbations or in incremental methods, one cannot exhibit a function which is guaranteed to decrease from one iteration of the algorithm to the next. This makes classical analysis not applicable. The key difference of the technique presented in this section is that the monotonicity requirement for $V(\cdot)$ is relaxed. We thus refer to $V(\cdot)$ as a pseudo-Lyapunov function. This generalization makes our approach applicable to a wider class of algorithms, including incremental methods and methods with perturbations.

We now state the generalized Lyapunov direct method. Convergence (attraction) properties of the process are expressed in terms of a pseudo-Lyapunov function $V(\cdot)$. For each specific algorithm, these properties allow further interpretation depending on the choice of $V(\cdot)$ for this algorithm.

We consider the following general iterative process:

$$x^{i+1} \in x^i - \eta_i G(i, x^i) - \xi^i, \qquad i=0, 1, \ldots, x^0 \in X', \xi^i \in \mathfrak{R}^n, \tag{6}$$

where

$$\lim_{i\to\infty} \eta_i = 0, \sum_{i=0}^{\infty} \eta_i = \infty, \quad \sum_{i=0}^{\infty} \xi^i \text{ is componentwise convergent}, \tag{7}$$

$G(\cdot,\cdot): \mathcal{N} \times X' \to \mathcal{M}(X')$, and $X'$ is an open set in $\mathfrak{R}^n$. In applications, $\xi^i$ usually corresponds to (random) noise. We further make a natural boundedness assumption on the mapping $G(\cdot,\cdot)$, in particular,

$$\sup_{x\in X'} \limsup_{\substack{x'\to x \\ i\to\infty}} \sup_{y\in G(i,x')} \|y\| < \infty.$$

Thus, the upper topological limit of $G(\cdot,\cdot)$, denoted by

$$G_0(x) := \operatorname*{\overline{lt}}_{\substack{x^i \to x \\ i \to \infty}} G(i, x^i),$$

is bounded and upper semicontinuous on a neighborhood of any compact set $X \subset X'$.

We assume that there exists a compact set $X \subset X'$ which contains all the accumulation points of the iterates generated by (6)–(7), that is,

$$\operatorname*{\overline{lt}}_{i \to \infty} \{x^i\} \subset X. \tag{8}$$

Let a pseudo-Lyapunov function $V(\cdot)$ be chosen (in applications, the choice depends on the problem and on the algorithm employed to solve it). Let $V(\cdot)$ be Lipschitz continuous and regular on a neighborhood of $X$. For the pseudo-Lyapunov function $V(\cdot)$, the set $X$, and the map $G_0(\cdot)$, we define the following set which is crucial for our analysis:

$$\mathscr{A}_0 := \left\{ x \in X \;\middle|\; \max_{h \in H(x)} \min_{g \in G_0(x)} \langle h, g \rangle \leq 0 \right\}, \tag{9}$$

where

$$H(x) = \operatorname{conv}\{\partial V(x) \cup N_X(x)\}.$$

Roughly speaking, the set $\mathscr{A}_0$ is comprised of all the points in $X$ for which $-G_0(x)$ does not contain feasible directions that are of descent for the pseudo-Lyapunov function $V(\cdot)$.

The following result shows that the sequences generated by (6)–(7) and satisfying (8) are, in a certain sense, attracted to the components of the set $\mathscr{A}_0$. We first have to introduce the notion of $V(\cdot)$-connected components of $\mathscr{A}_0$ (recall that $\mathscr{A}_0$ is compact). We say that a set $C \subset \mathfrak{R}^n$ is $V(\cdot)$-connected, if the set

$$V(C) = \{v \in \mathfrak{R} \mid \exists x \in C,\ v = V(x)\}$$

is a connected set in $\mathfrak{R}$. Let $\{\mathscr{A}^\gamma\}$, $\gamma \in \Gamma$ be the (unique) decomposition of $\mathscr{A}_0$ into $V(\cdot)$-connected components (see Ref. 25), that is

$$\mathscr{A}_0 = \bigcup_{\gamma \in \Gamma} \mathscr{A}^\gamma, \quad \mathscr{A}^{\gamma'} \neq \mathscr{A}^{\gamma''}, \qquad \text{for } \gamma' \neq \gamma'',\ \gamma', \gamma'' \in \Gamma.$$

The following theorem will play a central role in the subsequent analysis.

**Theorem 2.1.** See Ref. 3. For every sequence $\{x^i\}$ generated by the process (6)–(7), and satisfying (8), there exists a $\gamma \in \Gamma$ such that

$$\bar{\lim_{i\to\infty}} V(x^i) = V\left( \bar{\lim_{i\to\infty}} \{x^i\} \cap \mathscr{A}^\gamma \right).$$

Furthermore, every subsequence $\{x^{i_m}\}$ of $\{x^i\}$ satisfying

$$\lim_{m\to\infty} V(x^{i_m}) = \liminf_{i\to\infty} V(x^i) \quad \text{or} \quad \lim_{m\to\infty} V(x^{i_m}) = \limsup_{i\to\infty} V(x^i)$$

converges into $\mathscr{A}^\gamma$. In addition, if the set $V(\mathscr{A}_0)$ is nowhere dense in $\mathfrak{R}$, then the whole sequence $\{x^i\}$ converges into a connected component of $\mathscr{A}_0$.

We refer the reader to Ref. 3 for a detailed discussion.

## 3. Convergence Analysis

In this section, we consider a parallel generalized gradient-type projection method (GGPM) for solving the problem of minimizing an additive parametric objective function (2). The type of parallelization proposed here is primarily motivated by incremental gradient methods, particularly neural network training (Ref. 13). Empirical evaluation of parallel neural network training and numerical tests can be found in Ref. 26. We first consider the most general case. Our results can then be specialized by removing parallelism and/or considering the standard (nonadditive) objective function.

We now describe our notation for stating and establishing convergence properties of parallel GGPM in the presence of perturbations.

The index number $i = 1, 2, \ldots$ denotes major iterations of GGPM, each of which consists of going through the entire set of functions $f_1(x, \alpha_i), \ldots, f_K(x, \alpha_i)$. This is achieved in parallel by $p$ processors with processor $l$ handling at the $i$th iteration the functions $f_j(x, \alpha_i)$, $j \in J_l$, where $\{J_l\}$, $l = 1, \ldots, p$, is a partition of $\{1, \ldots, K\}$. Recall that $\alpha_i \in A$ is the (smoothing) parameter and $\lim_{i\to\infty} \alpha_i = \alpha_0$. For simplicity, we assume that the sets $\{J_l\}$ are ordered as follows:

$$J_1 = \{1, \ldots, K_1\},$$

$$J_2 = \{K_1 + 1, \ldots, K_1 + K_2\}, \ldots,$$

$$J_p = \{K_1 + \cdots + K_{p-1} + 1, \ldots, K\},$$

i.e.,

$$J_l=\{\bar{K}_l+1,\ldots,\bar{K}_l+K_l\},\qquad l=1,\ldots,p,$$
$$\text{where } \bar{K}_1=0,\qquad \bar{K}_l=\sum_{t=1}^{l-1}K_t,\ l=2,\ldots,p.$$

The index number $j=1,\ldots,K_l$ denotes minor iterations performed by the parallel processor $l$. Each minor iteration $j$ consists of a step in the direction of a negative generalized gradient $-\tilde{g}_l^{i,j}$ of the function $f_{\bar{K}_l+j}(\cdot,\alpha_i)$ at $z_l^{i,j}$, which is calculated with some error $\delta_l^{i,j}$,

$$\tilde{g}_l^{i,j}=g_l^{ij}+\delta_l^{i,j},\qquad g_l^{i,j}\in\partial f_{\bar{K}_l+j}(z_l^{i,j},\alpha_i),\qquad \delta_l^{i,j}=\delta_{\bar{K}_l+j}(z_l^{i,j},\alpha_i,i).$$

The function $\delta_j(z,\alpha,i)$ denotes a perturbation of the generalized gradient of $f_j(\cdot,\alpha)$ at the point $z\in X+\tau\mathbb{B}$ at the $i$th major iteration of the algorithm. Assuming that the perturbations are bounded, there exists a constant $M>0$ such that, for all $i$ and $j$,

$$\|y\|\le M,\qquad \forall y\in\partial f_j(x,\alpha_i)+\delta_j(x,\alpha_i,i),\qquad \forall x\in X+\tau\mathbb{B}. \tag{10}$$

The symbol $x^i$ refers to an iterate in $\mathfrak{R}^n$ of a major iteration $i=1,2,\ldots$. The symbol $z_l^{i,j}$ refers to an iterate in $\mathfrak{R}^n$ of a minor iteration $j=1,\ldots,K_l$, within a major iteration $i=1,2,\ldots$, computed by processor $l=1,\ldots,p$.

We consider the process with stepsizes decreasing subject to the following rule:

$$\lim_{i\to\infty}\eta_i=0,\qquad \sum_{i=0}^{\infty}\eta_i=\infty,\qquad \eta_i>0,\ i=0,1,\ldots. \tag{11}$$

We point out that the condition of stepsize going to zero is indispensable in the general nonsmooth case (see Ref. 19) as well as in the case of incremental methods (the latter is demonstrated in Ref. 27, Section 2).

We are now ready to state and prove the convergence properties of the parallel GGPM in the presence of perturbations.

**Algorithm 3.1.** Parallel GGPM. Start with any $x^0\in X$. Having $x^i$, compute $x^{i+1}$ as follows.

Step 1. Parallelization. For each processor $l\in\{1,\ldots,p\}$, compute

$$z_l^{i,j+1}=z_l^{i,j}-\eta_i\tilde{g}_l^{i,j},\qquad j=1,\ldots,K_l,\ \text{where } z_l^{i,1}=x^i. \tag{12}$$

Step 2. Synchronization. Compute

$$x^{i+1}=P_X\left[x^i+\sum_{l=1}^{p}(z_l^{i,K_l+1}-x^i)\right]. \tag{13}$$

Note that, for $K=1$, $p=1$, Algorithm 3.1 becomes a standard (perturbed) generalized gradient projection method, while $K\geq 2$, $p=1$ gives a serial incremental gradient-type method. Thus, the framework considered here is fairly general.

There are two sources of nonmonotonicity that come into play in Algorithm 3.1. First of all, each direction is associated with a generalized gradient of a partial objective function $f_j(\cdot, \alpha_l)$. Even if this direction is that of descent for $f_j(\cdot, \alpha_i)$, there is no guarantee that it is also of descent for the full objective function $f(\cdot, \alpha_0)$ given by (2) (also note a possible difference in the parameter value). The other source of nonmonotonicity is induced by perturbations of the generalized gradients.

We first note that it is easy to ensure that all the minor iterates remain within the set $X+\tau\mathbb{B}$, and hence are well defined. In particular, this holds if we choose

$$\eta_i \leq \tau/(M \max_l K_l), \tag{14}$$

where $M$ satisfies (10). We omit the proof, which is quite straightforward and is similar to Ref. 14, Lemma 2. From now on, we assume that the stepsizes satisfy both (11) and (14).

The following lemma will be used for translating Algorithm 3.1 into the framework of Section 2. We do not include the proof, which is fairly easy.

**Lemma 3.1.** Let $y=P_X[x-\eta g]$, where $x\in X$, $g\in\Re^n$, and $\eta>0$. Then, there exists $h\in N_X(y)$ such that

$$y=x-\eta(g+h) \quad \text{and} \quad \|h\|\leq\|g\|.$$

Using Lemma 3.1, we can rewrite the synchronization step (13) as follows:

$$x^{i+1}=x^i+\sum_{l=1}^{p}(z_l^{i,K_l+1}-x^i)+h, \qquad h\in N_X(x^{i+1}). \tag{15}$$

By (12), we have

$$z_l^{i,K_l+1}=x^i-\eta_i\sum_{j\in J_l}(g_l^{i,j}+\delta_l^{i,j}). \tag{16}$$

Combining (15) and (16), we obtain

$$x^{i+1}=x^i-\eta_i\sum_{l=1}^{p}\sum_{j\in J_l}(g_l^{i,j}+\delta_l^{i,j})+h,$$

where

$$h \in N_X(x^{i+1}), \qquad g_l^{i,j} \in \partial f_{\bar{K}_{l-1}+j}(z_l^{i,j+1}).$$

Using these relations, we next introduce a map $G(\cdot,\cdot): \mathcal{N} \times X \to \mathfrak{R}^n$ such that every sequence $\{x^i\}$ generated by Algorithm 3.1 is a trajectory of the iterative process

$$x^{i+1} \in x^i - \eta_i G(i, x^i), \qquad i=0, 1, \ldots, x^0 \in X.$$

We will refer to this $G(\cdot,\cdot)$ as the characteristic mapping of the algorithm:

$$G(i,x) = \left\{ v \in \mathfrak{R}^n \;\middle|\; \begin{array}{l} v = \sum_{l=1}^{p} \sum_{j \in J_l} (g_l^j + \delta_l^j) + h, \text{ where } h \in N_X(y),\ \|h\| \le MK, \text{ and} \\ y = x + \sum_{l=1}^{p} (z_l^{K_l+1} - x) + h,\ z_l^{j+1} = z_l^j - \eta_l (g_l^j + \delta_l^j),\ z_l^1 = x, \\ g_l^j \in \partial f_{\bar{K}_{l-1}+j}(z_l^j, \alpha_l),\ \delta_l^j = \delta_{\bar{K}_l+j}(z_l^j, \alpha_l, i),\ j=1, \ldots, K_l,\ l=1, \ldots, p \end{array} \right\}. \tag{17}$$

By Lemma 3.1 and (10), the map $G(\cdot,\cdot)$ is bounded, hence so is its upper topological limit.

To analyze the influence of computational errors $\delta_l^{i,j}$ on the convergence properties of the algorithm, we need to estimate the level of perturbations in the limit. We define $\epsilon(x)$, the exact asymptotic level of perturbations at a point $x \in X$, by the following relation:

$$\epsilon(x) = \limsup_{\substack{z_j (\in X + \epsilon \mathbb{B}) \to x (\in X) \\ i \to \infty}} \left\| \sum_{j=1}^{K} \delta_j(z_j, \alpha_i, i) \right\|. \tag{18}$$

It is easy to see that the function $\epsilon: X \to \mathfrak{R}_+$ is upper semicontinuous.

We are now ready to apply the generalized Lyapunov direct method of Section 2 to our algorithm. Our main result is the following theorem.

**Theorem 3.1.** For every sequence $\{x^i\}$ generated by Algorithm 3.1, there exists $X_s(\epsilon)^\gamma$, an $f(\cdot)$-connected component of $X_s(\epsilon)$, such that

$$\bar{\lim_{i \to \infty}} f(x^i) = f\left( \bar{\lim_{i \to \infty}} \{x^i\} \cap X_s(\epsilon)^\gamma \right).$$

Furthermore, every subsequence $\{x^{i_m}\}$ of $\{x^i\}$, satisfying

$$\lim_{m \to \infty} f(x^{i_m}) = \liminf_{i \to \infty} f(x^i) \text{ or } \lim_{m \to \infty} f(x^{i_m}) = \limsup_{i \to \infty} f(x^i), \tag{19}$$

converges into $X_s(\epsilon)^\gamma$. In addition, if $\epsilon(\cdot) \equiv 0$ and the set $f(X_s)$ is nowhere dense in $\mathfrak{R}$, then every sequence $\{x^i\}$ generated by Algorithm 3.1 converges into a connected component of $X_s$.

**Proof.** We choose $V(x) := f(x)$ as the pseudo-Lyapunov function of the iterative process. Following the approach outlined in Section 2, we introduce the set

$$\mathcal{A}_0 := \left\{ x \in X \;\middle|\; \max_{h \in H(x)} \min_{g \in G_0(x)} \langle h, g \rangle \leq 0 \right\},$$

where

$$H(x) := \operatorname{conv}\{\partial f(x) \cup N_X(x)\}.$$

Our proof is by virtue of showing that

$$\mathcal{A}_0 \subset X_s(\epsilon),$$

and then applying Theorem 2.1.

We first have to estimate $G_0(\cdot)$, the upper topological limit of $G(\cdot, \cdot)$. Because $\eta_i \to 0$, in (17) we have that $z_l^j \to x$, $j = 1, \ldots, K_l + 1$, $l = 1, \ldots, p$ and $y \to x$ as $x' \to x$, $i \to \infty$. Therefore, by the upper semicontinuity of $\partial f(\cdot, \cdot)$ and $N_X(\cdot)$ [see (5)] and the definition (18) of $\epsilon(\cdot)$, we have from (17) that

$$G_0(x) := \operatorname*{\bar{lt}}_{\substack{x' \to x \\ i \to \infty}} G(i, x') \subset \partial f(x) + N_X(x) + \epsilon(x)\mathbb{B}. \tag{20}$$

For every $x \in X$, we define

$$h_0(x) = \arg\min\{\|h\| \mid h \in \partial f(x) + N_X(x)\}.$$

Note that

$$\|h_0(x)\| = r(x);$$

see (3). Since $h_0(x)$ is the orthogonal projection of the origin onto the set $\{\partial f(x) + N_X(x)\}$, it follows that

$$\langle h_0(x), h \rangle \geq \|h_0(x)\|^2, \qquad \forall h \in \partial f(x) + N_X(x). \tag{21}$$

Since

$$h_0(x) \in \partial f(x) + N_X(x),$$

it follows that

$$(1/2)h_0(x) \in H(x). \tag{22}$$

Fix an arbitrary $x \notin X_s(\epsilon)$. By (4), we have

$$\|h_0(x)\| = r(x) > \epsilon(x). \tag{23}$$

We further obtain

$$\begin{aligned}
\max_{h\in H(x)} \min_{g\in G_0(x)} \langle h, g\rangle &\geq (1/2) \min_{g\in G_0(x)} \langle h_0(x), g\rangle \\
&\geq (1/2) \min_{g\in \partial f(x)+N_X(x)+\epsilon(x)\mathbb{B}} \langle h_0(x), g\rangle \\
&\geq (1/2) \min_{\delta\in\epsilon(x)\mathbb{B}} \min_{h\in\partial f(x)+N_X(x)} \langle h_0(x), h+\delta\rangle \\
&\geq (1/2) \min_{\delta\in\epsilon(x)\mathbb{B}} \langle h_0(x), h+\delta\rangle \\
&\geq (1/2) \min_{\delta\in\epsilon(x)\mathbb{B}} (\|h_0(x)\|^2 - \|\delta\|\,\|h_0(x)\|) \\
&\geq (1/2)\|h_0(x)\|(\|h_0(x)\| - \epsilon(x)) > 0,
\end{aligned}$$

where the first inequality follows from (22), the second inequality follows from (20), the fifth inequality follows from (21), and the last inequality follows from (23). Hence, $x\notin\mathcal{A}_0$, and it follows that $\mathcal{A}_0\subset X_s(\epsilon)$. Now, applying Theorem 2.1, we obtain immediately the desired results. □

We next consider a modification of the parallel GGPM, where a heavy ball term (Ref. 24) is added in the synchronization step,

$$x^{i+1} = P_X\left[x^i + \sum_{l=1}^{p} (z_l^{i,K_l+1} - x^i) + \beta_i(x^i - x^{i-1})\right]. \tag{24}$$

In neural network literature, methods of this type are usually referred to as backpropagation with momentum term (Refs. 8, 12). With respect to coefficients multiplying the heavy ball term, we assume that

$$\beta_i \geq 0, \quad i=0, 1, \ldots, \qquad \lim_{i\to\infty} \beta_i = 0. \tag{25}$$

We also make the following mild assumption on the stepsizes, in addition to (11), (14):

$$\limsup_{i\to\infty} (\eta_{i-1}/\eta_i) < +\infty. \tag{26}$$

The next result shows that methods with a heavy ball term possess the same convergence and stability properties as the gradient projection methods.

**Theorem 3.2.** Let $\{x^i\}$ be a sequence generated by the parallel GGPM with a heavy ball synchronization step (24). Then, all the conclusions of Theorem 3.1 hold.

**Proof.** We show that the upper topological limits of the characteristic mappings for GGPM with and without a heavy ball term are essentially the

same (note that the mappings themselves are certainly different). We first define the following quantity:

$$\mu_i := 2\beta_i KM(\eta_{i-1}/\eta_i), \qquad i=1,2,\ldots, \qquad \mu_0=0.$$

By (25) and (26),

$$\lim_{i\to\infty} \mu_i = 0.$$

By the construction of the algorithm and (10),

$$\beta_i(x^i - x^{i-1}) \in x^i + 2\beta_i KM\eta_{i-1}\mathbb{B} = x^i + \eta_i\mu_i\mathbb{B}.$$

Let us denote the characteristic map of GGPM with a heavy ball term by $\tilde{G}(\cdot,\cdot)$. Then, we have that

$$\tilde{G}(i,x^i) \subset G(i,x^i) + \mu_i\mathbb{B},$$

where $G(\cdot,\cdot)$ is the characteristic map of Algorithm 3.1 defined by (17). Therefore,

$$x^{i+1} \in x^i - \eta_i\tilde{G}(i,x^i) \subset x^i - \eta_i(G(i,x^i) + \mu_i\mathbb{B}).$$

Now, taking into account that $\mu_i \to 0$, we obtain

$$\begin{aligned}\tilde{G}_0(x) &:= \operatorname*{\overline{lt}}_{\substack{x'\to x\\ i\to\infty}} \tilde{G}(i,x^i) \subset \operatorname*{\overline{lt}}_{\substack{x'\to x\\ i\to\infty}} (G(i,x') + \mu_i\mathbb{B}) \\ &= \operatorname*{\overline{lt}}_{\substack{x'\to x\\ i\to\infty}} G(i,x') = G_0(x).\end{aligned}$$

Hence, by (20),

$$\tilde{G}_0(x) \subset \partial f(x) + N_X(x) + \epsilon(x)\mathbb{B}.$$

The rest of the proof is analogous to that of Theorem 3.1, and thus is omitted. □

## 4. Important Special Cases

In this section, we consider the standard optimization problem (1) of minimizing a Lipschitz continuous regular function over a convex compact set, and establish stronger convergence properties of the generalized subgradient projection method in a number of important special cases. These include problems with relatively small perturbations, convex and strongly convex problems, and problems with weak sharp minima (Ref. 28).

We start with the following lemma, which deals with the case of perturbations small relative to the residual function $r(\cdot)$ defined in (3).

**Lemma 4.1.** Let $\epsilon(x) \leq \max\{\bar{\epsilon}, \lambda r(x)\}, \forall x \in X$, where $\bar{\epsilon} \geq 0, \lambda \in [0, 1)$. Then, $X_s(\epsilon) \subset X_s(\bar{\epsilon})$. In particular, if $\bar{\epsilon} = 0$, then $X_s(\epsilon) = X_s$.

**Proof.** Suppose that $x \in X_s(\epsilon)$. Then, by (4) and the assumption of the lemma, we have

$$r(x) \leq \epsilon(x) \leq \max\{\bar{\epsilon}, \lambda r(x)\}.$$

If $\lambda r(x) \geq \bar{\epsilon}$, then $r(x) \leq \lambda r(x)$ and $1 > \lambda \geq 0$ imply that $r(x) = 0$. Since $X_x(0) \subset X_s(\bar{\epsilon})$, we have that $x \in X_s(\bar{\epsilon})$. If $\lambda r(x) \leq \bar{\epsilon}$, then $r(x) \leq \epsilon(x) \leq \bar{\epsilon}$, and hence $x \in X_s(\bar{\epsilon})$. □

Let $d(\cdot, C)$ be the distance function to the set $C \subset \mathfrak{R}^n$, that is,

$$d(x, C) = \inf_{y \in C} \|x - y\|.$$

Define

$$\bar{\epsilon} = \sup_{x \in X} \epsilon(x), \qquad D = \sup_{x, y \in X} \|x - y\|.$$

The following lemma relates the $\epsilon$-stationary sets to the $\epsilon$-optimal sets for the case where $f(\cdot)$ is convex. This result was also used in Ref. 29.

**Lemma 4.2.** Let $f(\cdot)$ be convex on $X$. Then, $X_s(\epsilon(x)) \subset X_{\text{opt}}(\epsilon(x) d(x, X_{\text{opt}}))$. In particular, $X_s(\epsilon) \subset X_{\text{opt}}(\bar{\epsilon} D)$. In addition, if $f(\cdot)$ is differentiable and strongly convex on $X$ with modulus $\theta > 0$, and if $X_s(\bar{\epsilon}) \subset \text{int } X$, then $X_s(\bar{\epsilon}) \subset X_{\text{opt}}(\bar{\epsilon}^2 / 2\theta)$.

**Proof.** Let $x \in X_s(\epsilon(x))$. By definition of $X_s(\epsilon)$, there exist $g \in \partial f(x)$, $h_1 \in N_X(x)$, and $h_2 \in \epsilon(x)\mathbb{B}$ such that

$$0 = g + h_1 + h_2.$$

Let $x^*$ be the closest point to $x$ in $X_{\text{opt}}$. By convexity of $f(\cdot)$, it follows that

$$\begin{aligned} f(x) - f(x^*) &\leq \langle -g, x^* - x \rangle = \langle h_1 + h_2, x^* - x \rangle \\ &\leq \langle h_2, x^* - x \rangle \leq \epsilon(x) d(x, X_{\text{opt}}), \end{aligned}$$

where the second inequality follows from the definition of the normal cone. This establishes the first two assertions of the lemma.

For the last assertion, just note that (see Ref. 24, p. 24), for any $x \in X$,

$$2\theta(f(x) - \min_{y \in X} f(y)) \leq \|\partial f(x)\|^2.$$ □

Recall that $X_{\text{opt}}$ is a set of weak sharp minima (Ref. 28) with parameter $\rho>0$ if

$$f(x)-\min_{y\in X} f(y)\geq \rho d(x, X_{\text{opt}}), \qquad \forall x\in X.$$

The following lemma shows that, for problems with weak sharp minima, $\epsilon$-stationary sets coincide with the set of minima, provided $\epsilon$ is small relative to the parameter $\rho$.

**Lemma 4.3.** Let $f(\cdot)$ be convex on $X$. Assume that $X_{\text{opt}}$ is a set of weak sharp minima with parameter $\rho>0$. Then, if $\epsilon(x)\leq \max\{\nu, \lambda r(x)\}$, $\forall x\in X$, where $\lambda\in[0, 1)$ and $\nu\in[0, \rho)$, it follows that $X_s(\epsilon)=X_{\text{opt}}$.

**Proof.** Obviously, $X_{\text{opt}}\subset X_s(\epsilon)$. Take any $x\in X_s(\epsilon)$. By Lemmas 4.1 and 4.2, and our assumption, we have

$$x\in X_s(\epsilon(\cdot))\subset X_s(\nu)\subset X_{\text{opt}}(\nu d(x, X_{\text{opt}})).$$

Hence,

$$\nu d(x, X_{\text{opt}})\geq f(x)-\min_{y\in X} f(y)\geq \rho d(x, X_{\text{opt}}).$$

Now, $\nu<\rho$ implies that $d(x, X_{\text{opt}})=0$, that is $x\in X_{\text{opt}}$. □

In conclusion, we summarize the convergence and stability properties of the serial generalized gradient projection method with a heavy ball term.

**Algorithm 4.1.** GGPM with a Heavy Ball Term. Start with any $x^0\in X$. Having $x^i$, compute $x^{i+1}$ as follows:

$$x^{i+1}=P_X[x^i-\eta_i(g_i+\delta(x^i, \alpha_i, i))+\beta_i(x^i-x^{i-1})],$$

$$g_i\in\partial f(x^i, \alpha_i), \qquad i=0, 1, \ldots,$$

where the parameters $\eta_i$, $\alpha_i$, $\beta_i$ are the same as specified in Section 3.

Combining Theorems 3.1, 3.2, and Lemmas 4.1 to 4.3, we obtain immediately the following convergence results for Algorithm 4.1.

**Theorem 4.1.** Every sequence $\{x^i\}$ generated by Algorithm 4.1 possesses the following properties:

(i) there exists an $f(\cdot)$-connected component $X_s(\epsilon)^\gamma$ of $X_s(\epsilon)$ such that

$$\bar{\lim_{i\to\infty}}\{f(x^i)\}=f(\bar{\lim}\{x^i\}\cap X_s(\epsilon)^\gamma);$$

(ii) every subsequence $\{x^{i_m}\}$ of $\{x^i\}$ satisfying (19) converges into $X_s(\epsilon)^\gamma$;

(iii) if the perturbations are relatively small, that is, $\epsilon(x) \leq \lambda r(x)$, for all $x \in X$ and some $\lambda \in [0, 1)$, and if the set $f(X_s)$ is nowhere dense in $\mathfrak{R}$, then $\{x^i\}$ converges into $X_s$;

(iv) if $f(\cdot)$ is convex, then $\{x^i\}$ converges into the set

$$X_{\text{opt}}(\epsilon(x)d(x, X_{\text{opt}})) \subset X_{\text{opt}}(\bar{\epsilon}D);$$

(v) if $f(\cdot)$ is convex, $X_{\text{opt}}$ is a set of weak sharp minima with parameter $\rho > 0$, and if

$$\epsilon(x) < \rho, \qquad \forall x \in X,$$

then $\{x^i\}$ converges into $X_{\text{opt}}$;

(vi) if $f(\cdot)$ is strongly convex with modulus $\theta > 0$, and if

$$X_s(\bar{\epsilon}) \subset \text{int } X, \qquad \text{where } \bar{\epsilon} := \sup_{x \in X} \epsilon(x),$$

then $\{x^i\}$ converges into $X_{\text{opt}}(\bar{\epsilon}^2/2\theta)$.

Theorem 4.1 extends, strengthens, and unifies results on convergence and stability properties of the generalized subgradient projection method given in Refs. 19–21.

## References


1. CLARKE, F. H., *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, New York, 1983.
2. ALBER, YA. I., IUSEM, A. N., and SOLODOV, M. V., *On the Projected Subgradient Methods for Nonsmooth Convex Optimization in a Hilbert Space*, Mathematical Programming, Vol. 81, pp. 23–25, 1998.
3. ZAVRIEV, S. K., and PEREVOZCHIKOV, A. G., *Direct Lyapunov Method in Attraction Analysis of Finite-Difference Inclusions*, USSR Computational Mathematics and Mathematical Physics, Vol. 30, pp. 22–32, 1990.
4. ZAVRIEV, S. K., *Convergence Properties of the Gradient Method under Variable Level Interference*, USSR Computational Mathematics and Mathematical Physics, Vol. 30, pp. 997–1007, 1990.
5. SOLODOV, M. V., *Convergence Analysis of Perturbed Feasible-Descent Methods*, Journal of Optimization Theory and Applications, Vol. 92, pp. 337–353, 1997.
6. SOLODOV, M. V., and SVAITER, B. F., *Descent Methods with Line Search in the Presence of Perturbations*, Journal of Computational and Applied Mathematics, Vol. 80, pp. 265–275, 1997.
7. BERTSEKAS, D. P., *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1995.

8. KHANNA, T., *Foundations of Neural Networks*, Addison-Wesley, Reading, Massachusetts, 1989.
9. ANDERSON, B. D. O., and MOORE, J. B., *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, New Jersey, 1979.
10. MAYNE, D. Q., and POLAK, E., *Nondifferentiable Optimization via Adaptive Smoothing*, Journal of Optimization Theory and Applications, Vol. 43, pp. 601–614, 1984.
11. MANGASARIAN, O. L., *Mathematical Programming in Neural Networks*, ORSA Journal on Computing, Vol. 5, pp. 349–360, 1993.
12. MANGASARIAN, O. L., and SOLODOV, M. V., *Backpropagation Convergence via Deterministic Nonmonotone Perturbed Minimization*, Neural Information Processing Systems, Edited by G. Tesauro, J. D. Cowan, and J. Alspector, Morgan Kaufmann Publishers, San Francisco, California, Vol. 6, pp. 383–390, 1994.
13. MANGASARIAN, O. L., and SOLODOV, M. V., *Serial and Parallel Backpropagation Convergence via Nonmonotone Perturbed Minimization*, Optimization Methods and Software, Vol. 4, pp. 103–116, 1994.
14. LUO, Z. Q., and TSENG, P., *Analysis of an Approximate Gradient-Projection Method with Applications to the Backpropagation Algorithm*, Optimization Methods and Software, Vol. 4, pp. 85–101, 1994.
15. BERTSEKAS, D. P., *Incremental Least-Squares Methods and the Extended Kalman Filter*, SIAM Journal on Optimization, Vol. 6, pp. 807–822, 1996.
16. SOLODOV, M. V., *Incremental Gradient Algorithms with Stepsizes Bounded Away from Zero*, Computational Optimization and Applications (to appear).
17. BERTSEKAS, D. P., *A New Class of Incremental Gradient Methods for Least-Squares Problems*, SIAM Journal on Optimization, Vol. 7, pp. 913–926, 1997.
18. TSENG, P., *Incremental Gradient-Projection Method with Momentum Term and Adaptive Stepsize Rule*, SIAM Journal on Optimization, Vol. 8, pp. 506–531, 1998.
19. MIKHALEVITCH, V. S., GUPAL, A. M., and NORKIN, V. I., *Methods of Nonconvex Optimization*, Nauka, Moscow, Russia, 1987 (in Russian).
20. DOROFEEV, P. A., *On Some Properties of the Quasi-Gradient Method*, USSR Computational Mathematics and Mathematical Physics, Vol. 25, pp. 181–189, 1985.
21. ZAVRIEV, S. K., and PEREVOZCHIKOV, A. G., *Attraction of Trajectories of Finite-Difference Inclusions and Stability of Numerical Methods of Stochastic Nonsmooth Optimization*, Soviet Physics Doklady, Vol. 313, pp. 1373–1376, 1990.
22. ROUCHE, N., HABETS, P., and LALOY, M., *Stability Theory by Liapunov Direct Method*, Springer Verlag, New York, New York, 1977.
23. ZANGWILL, W. I., *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
24. POLYAK, B. T., *Introduction to Optimization*, Optimization Software, Publications Division, New York, New York, 1987.
25. ZAVRIEV, S. K., *Stochastic Subgradient Methods for Minmax Problems*, Izdatelstvo MGU, Moscow, Russia, 1984 (in Russian).

26. PAUGAM-MOISY, H., *On Parallel Algorithm for Backpropagation by Partitioning the Training Set*, Proceedings of the 5th International Conference on Neural Networks and Their Applications, Nimes, France, 1992.
27. LUO, Z. Q., *On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks*, Neural Computation, Vol. 3, pp. 226–245, 1991.
28. BURKE, J. V., and FERRIS, M. C., *Weak Sharp Minima in Mathematical Programming*, SIAM Journal on Control and Optimization, Vol. 31, pp. 1340–1359, 1993.
29. SOLODOV, M. V., *New Inexact Parallel Variable Distribution Algorithms*, Computational Optimization and Applications, Vol. 7, pp. 165–182, 1997.