JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS Vol. 93, No 2, pp. 337-353, MAY 1997

Convergence Analysis of Perturbed Feasible Descent Methods[1]

M. V. Solodov[2]

Communicated by Z. Q. Luo

Abstract. We develop a general approach to convergence analysis of feasible descent methods in the presence of perturbations. The important novel feature of our analysis is that perturbations need not tend to zero in the limit. In that case, standard convergence analysis techniques are not applicable. Therefore, a new approach is needed. We show that, in the presence of perturbations, a certain ϵ-approximate solution can be obtained, where ϵ depends linearly on the level of perturbations. Applications to the gradient projection, proximal minimization, extragradient and incremental gradient algorithms are described.

Key Words. Feasible descent methods, perturbation analysis, approximate solutions.

1. Introduction

We consider the general mathematical programming problem of minimizing a differentiable function f: $\Re^n \to \Re$ over a closed convex set X in $\Re^n$,

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

We assume that $f\in C_L^1(X)$, that is, $f(\cdot)$ has Lipschitz continuous partial derivatives on X,

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|, \qquad \forall x\in X, y\in X, \tag{2}$$

where L is a positive scalar, $\nabla f(\cdot)$ denotes the gradient of $f(\cdot)$, and $\|\cdot\|$ denotes the Euclidean norm.

[1] The author was supported in part by CNPq Grant 300734/95-6. This work was started when the author was with the Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, where he was supported by Air Force Office of Scientific Research Grant F49620-94-1-0036 and National Science Foundation Grant CCR-9322479.

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

Let $[\cdot]^+$ denote the orthogonal projection onto X. Following Ref. 1, we consider a broad class of feasible descent methods that can be represented by the formula

$$x^{\text{new}} := [x - \eta \nabla f(x) + e(x, \eta)]^+, \tag{3}$$

where η is a positive scalar and the mapping $e: \Re^{n+1} \to \Re^n$ is the defining feature of each particular algorithm; see Section 3. This is a rather general framework that includes the gradient projection algorithm (Refs. 2 and 3), proximal minimization algorithm (Refs. 4 and 5), extragradient algorithm (Refs. 6 and 7), and incremental gradient algorithms (Ref. 8), among others. We note in passing that, in the noise-free case, the characteristic mappings $e(\cdot, \cdot)$ of classical feasible descent methods satisfy the condition $e(x^i, \eta_i) \to 0$ as $i \to \infty$ by the algorithm construction; see Ref. 1 for details. Only incremental methods are an exception to this rule; see Ref. 8 for details.

In this paper, we are concerned with the behavior of feasible descent algorithms in the presence of perturbations,

$$x^{\text{new}} := [x - \eta \nabla f(x) + e(x, \eta) + \delta(x, \eta)]^+. \tag{4}$$

Here, $e(\cdot, \cdot)$ plays the same role as in (3), namely, it is the characteristic of the method, while $\delta(\cdot, \cdot)$ represents perturbations due to inexact computation of the gradient of $f(\cdot)$, or inexact subproblem solution, or both. We say that perturbations are essential (nonvanishing) if

$$\delta(x^i, \eta_i) \not\to 0, \qquad \text{as } i \to \infty.$$

In this paper, we consider nonvanishing perturbations and make only the mild assumption that the perturbations are uniformly bounded,

$$\|\delta(x, \eta)\| \le \bar{\epsilon}, \qquad \text{for some } \bar{\epsilon} > 0,\ \forall x \in X,\ \eta \le \bar{\eta}. \tag{5}$$

The latter is the only practical assumption in the case where the perturbations cannot be effectively controlled. This may happen, for example, when the function and/or gradient values are not given explicitly, but instead are computed as an approximate solution of some possibly difficult subproblem. We note that very little is known about the convergence properties of essentially perturbed algorithms. The primary contribution of this paper is laying down theoretical framework for analysis of such algorithms.

Convergence and rate of convergence of feasible descent methods have been studied extensively; see Ref. 1 and references therein. We point out that the previous work either deals with the case where no perturbations are present $[\delta(x^i, \eta_i) = 0]$ or assumes some conditions that explicitly or implicitly imply that perturbations vanish in the limit $[\delta(x^i, \eta_i) \to 0]$. Some conditions of this type have been used in the analysis of matrix splitting methods,

$$\|\delta(x^i, \eta_i)\| \leq c\|x^{i+1} - x^i\|, \qquad c > 0,\ c \text{ sufficiently small},$$

or

$$\sum_{i=0}^{\infty} \|\delta(x^i, \eta_i)\| < \infty.$$

In particular, the first condition has been used in Refs. 9 and 10 and the second in Ref. 11. Note that either assumption ensures that

$$\delta(x^i, \eta_i) \to 0, \qquad \text{as } i \to \infty.$$

Similar tolerance requirements are common in other methods that involve solving subproblems (e.g., Refs. 12 and 13). In the above-mentioned cases, the convergence properties of the algorithm stay intact, except possibly for the rate of convergence. We emphasize that the setting considered in this work is fundamentally different. Condition (5) no longer guarantees the convergence of the iterates generated by (4) to an exact solution of (1). Moreover, standard relations such as

$$f(x^i) - f(x^{i+1}) \geq 0,$$

$$\|x^{i+1} - x^i\| \to 0, \qquad \text{as } i \to \infty,$$

need not hold; see Section 2. This makes traditional convergence analysis techniques (Refs. 14 and 15) inapplicable. In this paper, we develop a new approach to the analysis of feasible descent algorithms with nonvanishing perturbations. Our analysis extends some of the ideas presented in Refs. 16–19 for methods of unconstrained optimization and is close in spirit to the study of an unconstrained gradient method in Ref. 19. Essential perturbations were considered in Ref. 20 in a different context of incremental gradient-type methods with decaying stepsize. A special case of an approximate gradient projection method with decaying stepsize is also studied in Ref. 21. We note that, in the present paper, the stepsize is bounded away from zero. Therefore, the situation and the analysis required are completely different from those of Refs. 20 and 21.

We now define the following residual function:

$$r(x) := x - [x - \nabla f(x)]^+,$$

which is central for the subsequent analysis. It is well known that some $\bar{x} \in \Re^n$ satisfies the minimum principle optimality condition (Ref. 22) for problem (1) if and only if $r(\bar{x}) = 0$. We shall call such $\bar{x}$ a stationary point of (1). For a nonnegative upper semicontinuous function $\epsilon: \Re^n \to \Re_+$, we define an $\epsilon(\cdot)$-stationary set of problem (1) as follows:

$$S(\epsilon(\cdot)) := \{x \in X \mid \|r(x)\| \leq \epsilon(x)\}. \tag{6}$$

Clearly, $S(0)$ is the set of all stationary points in the usual sense [we shall use the notation $S := S(0)$]. In Section 2, we show that, for any bounded sequence of iterates generated by (4), there exists at least one accumulation point which is in the set $S(\epsilon)$, with ϵ depending linearly on the level of perturbations.

We note that another important property of the residual function $r(\cdot)$ is that, under certain conditions, its norm provides a (local) upper bound on the distance to the set S; see Refs. 1 and 23. Namely, there exist positive constants μ and ν, depending on $f(\cdot)$ and X only, such that

$$d(x, S) \leq \mu \|r(x)\|, \qquad \forall x \text{ with } \|r(x)\| \leq \nu, \tag{7}$$

where $d(\cdot, S)$ denotes the Euclidean distance to S. Moreover, under additional assumptions, this condition holds with $\nu = \infty$ (global error bound, Refs. 24–26). Therefore, if $x \in S(\epsilon)$ and the bound (7) holds with $\nu \geq \epsilon$, it follows immediately that

$$d(x, S) \leq \mu \|r(x)\| \leq \mu \epsilon.$$

The rest of the paper is organized as follows. In Section 2, we develop our general technique for convergence analysis of perturbed algorithms. In Section 3, we show how our results apply to the gradient projection, proximal point, extragradient and incremental gradient algorithms. Section 4 contains some concluding remarks.

One more word about our notation. The usual inner product of two vectors $x \in \Re^n$, $y \in \Re^n$ is noted by $\langle x, y \rangle$. The Euclidean 2-norm of $x \in \Re^n$ is given by $\|x\|^2 = \langle x, x \rangle$. For a bounded sequence $\{x^i\}$ in $\Re^n$, $\mathrm{lt}_{i \to \infty}\{x^i\}$ denotes the set of all accumulation points of $\{x^i\}$. For two nonnegative scalar functions $s_1 : \Re_+ \to \Re_+$ and $s_2 : \Re_+ \to \Re_+$, we say that $s_1 = O(s_2)$ if there exists a positive constant c such that

$$\lim_{t \to \infty} s_1(t)/s_2(t) = c.$$

2. Convergence Analysis of Methods with Perturbations

In this section, we present our general framework for the analysis of feasible descent methods in the presence of essential perturbations. Our argument is based on monitoring the behavior of $f(\cdot)$ on the iterates of the algorithm. We emphasize that this behavior is nonmonotone and Lyapunov-type convergence analysis (Refs. 15 and 27) cannot be applied.

We first state three well-known results that will be used later.

Lemma 2.1. See Ref. 14, p.6. Let $f(\cdot)\in C_L^1(X)$. Then,

$$|f(y)-f(x)-\langle\nabla f(x), y-x\rangle|\leq(L/2)\|y-x\|^2, \qquad \forall x, y\in X.$$

Lemma 2.2. See Ref. 14, p.121. For any $x\in\Re^n$, any $y\in\Re^n$, and any $z\in X$, the following relations hold:

$$\langle y-[y]^+, z-[y]^+\rangle\leq 0,$$
$$\|[x]^+-[y]^+\|\leq\|x-y\|.$$

Lemma 2.3. See Ref. 28, Lemma 1. For any $x\in\Re^n$, any $y\in\Re^n$, and any $\eta>0$,

$$\begin{aligned}&\max\{1, \eta\}\|x-[x-y]^+\|\\ &\geq\|x-[x-\eta y]^+\|\\ &\geq\min\{1, \eta\}\|x-[x-y]^+\|.\end{aligned}$$

The method under consideration is the following model algorithm.

Algorithm 2.1. Start with any $x^0\in X$. For $i=0, 1, 2, \ldots,$ let

$$x^{i+1}\in T(x^i, \eta_i),$$

where

$$T(x, \eta)=[x-\eta\nabla f(x)+e(x, \eta)+\delta(x, \eta)]^+,$$

and the following conditions are satisfied:

$$\|e(x, \eta)\|\leq c_1\|x-T(x, \eta)\|, \qquad 0\leq c_1<1, \tag{8}$$

$$\langle e(x, \eta), x-T(x, \eta)\rangle\geq -c_2\|x-T(x, \eta)\|^2, \qquad 0\leq c_2<1, \tag{9}$$

$$c_3\leq\liminf_i \eta_i, \qquad \limsup_i \eta_i\leq\min\{1, 2(1-c_2)/L-c_3\}, \tag{10}$$

where

$$0<c_3<(1-c_2)/L.$$

In Section 3, we show that various important optimization methods fall within the framework of Algorithm 2.1. Condition (8) is standard for feasible descent methods and is a consequence of the algorithm construction (Ref. 1). The bounds (10) imposed on the stepsize are also fairly standard. With respect to (9), we note the following. If the left-hand-side of (9) is nonnegative for all x, then we set $c_2 := 0$; otherwise, we set $c_2 := c_1$. It follows that $0 \leq c_2 < 1$.

To study the convergence properties of Algorithm 2.1, we need to estimate the level of perturbations in the limit. We say that $\epsilon(x)$ is the exact asymptotic level of perturbations (Ref. 20) at a point $x \in X$, if

$$\epsilon(x) = \limsup_{\substack{y^k (\in X) \to x \\ k \to \infty}} \|\delta(y^k, \eta_k)\|.$$

It is easy to see that $\epsilon(\cdot): \Re^n \to \Re_+$ is upper semicontinuous.

For clarity of presentation, we briefly outline our argument. Using Lemmas 2.1–2.3 and conditions (8)–(10), we show that

$$f(x) - f(T(x, \eta)) \geq \varphi(x),$$

where $\varphi(x)$ is a certain lower semicontinuous function which depends on the residual $r(x)$ and the asymptotic level of perturbations $\epsilon(x)$; note that $\varphi(\cdot)$ need not be nonnegative. If $f(\cdot)$ is bounded from below on X, then for any sequence of iterates generated by Algorithm 2.1, there must exist at least one accumulation point belonging to the level set $\{x \in X \mid \varphi(x) \leq 0\}$; otherwise, we get a contradiction. Finally, using the dependence of $\varphi(\cdot)$ on $r(\cdot)$ and $\varepsilon(\cdot)$, we establish a certain relationship between the level sets of $\varphi(\cdot)$ and the $\epsilon(\cdot)$-stationary sets (6) of problem (1).

We are now ready to state and prove our main result.

Theorem 2.1. Suppose that $f \in C_L^1(X)$ and $f(\cdot)$ is bounded from below on X. Let conditions (8)–(10) be satisfied. Then, there exist positive constants d_1 and d_2 such that:

(i) for every bounded sequence $\{x^i\}$ generated by Algorithm 2.1, there exists an accumulation point $\bar{x}$ of $\{x^i\}$ such that

$$\bar{x} \in S(d_1 \epsilon(\cdot));$$

(ii) for every subsequence $\{x^{i_m}\}$ of $\{x^i\}$ satisfying

$$\limsup_{m \to \infty} f(x^{i_m}) \leq \liminf_{i \to \infty} f(x^i) + t, \qquad \text{for some } t \geq 0,$$

it follows that

$$\overline{\mathrm{lt}}_{m \to \infty} \{x^{i_m}\} \subset S(d_1 \epsilon(\cdot) + d_2 t^{1/2});$$

(iii) in particular, if the sequence $\{f(x^i)\}$ converges, then

$$\overline{\mathrm{lt}}_{i \to \infty} \{x^i\} \subset S(d_1 \epsilon(\cdot)).$$

Proof. Let $x:=x^i$ and $\eta:=\eta_i$. Then, for every $i=0, 1, 2, \ldots,$ by Lemma 2.1,

$$f(x)-f(T(x, \eta))\geq -\langle \nabla f(x), T(x, \eta)-x\rangle-(L/2)\|T(x, \eta)-x\|^2. \tag{11}$$

By Lemma 2.2, taking

$$y=x-\eta\nabla f(x)+e(x, \eta)+\delta(x, \eta) \quad \text{and} \quad z=x\in X,$$

we have

$$\langle x-\eta\nabla f(x)+e(x, \eta)+\delta(x, \eta)-T(x, \eta), x-T(x, \eta)\rangle\leq 0.$$

Hence,

$$\begin{aligned}&-\langle \nabla f(x), T(x, \eta)-x\rangle\geq(1/\eta)\\ &\times[\|x-T(x, \eta)\|^2+\langle e(x, \eta)+\delta(x, \eta), x-T(x, \eta)\rangle].\end{aligned}$$

Using (9), we have

$$\begin{aligned}&-\langle \nabla f(x), T(x, \eta)-x\rangle\geq(1/\eta)\\ &\times[(1-c_2)\|x-T(x, \eta)\|^2+\langle \delta(x, \eta), x-T(x, \eta)\rangle].\end{aligned}$$

Combining the latter inequality with (11), we further obtain

$$\begin{aligned}&f(x)-f(T(x, \eta))\\ &\geq[(1-c_2)/\eta-L/2]\|T(x, \eta)-x\|^2+(1/\eta)\langle \delta(x, \eta), x-T(x, \eta)\rangle\\ &\geq[(1-c_2)/\eta-L/2]\|T(x, \eta)-x\|^2+(1/\eta)\|\delta(x, \eta)\|\|x-T(x, \eta)\|\\ &\geq(1/\eta)(1-c_2-L\eta/2)\|T(x, \eta)-x\|^2-(1/\eta)\epsilon(x)\|T(x, \eta)-x\|\\ &\geq\{L^2c_3/[4(1-c_2)-2Lc_3]\}\|T(x, \eta)-x\|^2-(1/c_3)\epsilon(x)\|T(x, \eta)-x\|, \end{aligned} \tag{12}$$

where the second relation follows from the Cauchy-Schwartz inequality, the third inequality follows from the definition of $\epsilon(\cdot)$, and the last inequality follows from (10) for i sufficiently large, say $i\geq i_1$.

By Lemmas 2.3 and 2.2, the triangle inequality, and (8), it follows that

$$\begin{aligned}&\min\{1, \eta\}\|r(x)\|\\ &\leq\|x-[x-\eta\nabla f(x)]^+\|\\ &\leq\|x-T(x, \eta)\|+\|T(x, \eta)-[x-\eta\nabla f(x)]^+\|\\ &\leq\|x-T(x, \eta)\|+\|e(x, \eta)+\delta(x, \eta)\|\\ &\leq(1+c_1)\|x-T(x, \eta)\|+\epsilon(x).\end{aligned}$$

For $i \geq i_1$, using (10), we obtain

$$\|x - T(x, \eta)\| \geq [1/(1+c_1)][c_3\|r(x)\| - \epsilon(x)]. \tag{13}$$

Similarly,

$$\begin{aligned} &\|x - T(x, \eta)\| \\ &= \|x - [x - \eta\nabla f(x)]^+ + [x - \eta\nabla f(x)]^+ - T(x, \eta)\| \\ &\leq \max\{1, \eta\}\|r(x)\| + \|e(x, \eta) + \delta(x, \eta)\| \\ &\leq \|r(x)\| + c_1\|x - T(x, \eta)\| + \epsilon(x). \end{aligned}$$

Hence,

$$\|x - T(x, \eta)\| \leq [1/(1-c_1)][\|r(x)\| + \epsilon(x)]. \tag{14}$$

For $i \geq i_1$, combining (12)-(14) yields

$$\begin{aligned} f(x) - f(T(x, \eta)) \geq &\{L^2c_3/2(1+c_1)^2[2(1-c_2) - Lc_3]\}[c_3\|r(x)\| - \epsilon(x)]^2 \\ &- [1/c_3(1-c_1)]\epsilon(x)[\|r(x)\| + \epsilon(x)] \\ &= b_1\|r(x)\|^2 - b_2\epsilon(x)\|r(x)\| - b_3\epsilon(x)^2, \end{aligned}$$

where

$$\begin{aligned} b_1 &:= L^2c_3^3/2(1+c_1)^2[2(1-c_2) - Lc_3], \\ b_2 &:= L^2c_3^2/2(1+c_1)^2[2(1-c_2) - Lc_3] + 1/c_3(1-c_1), \\ b_3 &:= 1/c_3(1-c_1) - L^2c_3/2(1+c_1)^2[2(1-c_2) - Lc_3]. \end{aligned}$$

By (8)-(10), it is easy to see that $b_1 > 0$ and $b_2 > 0$. We next check that $b_3 > 0$. By (10),

$$1/c_3 \geq L/[2(1-c_2) - Lc_3].$$

Hence,

$$L^2c_3/2(1+c_1)^2[2(1-c_2) - Lc_3] \leq L/2 < (1-c_2)/c_3 < 1/c_3(1-c_1),$$

where the second inequality follows from (10). Hence, $b_3 > 0$.

We next define the following auxiliary function $\varphi: X \to \Re$, which is crucial for our analysis:

$$\varphi(x) := b_1\|r(x)\|^2 - b_2\epsilon(x)\|r(x)\| - b_3\epsilon(x)^2.$$

With this definition, we have

$$f(x) - f(T(x, \eta)) \geq \varphi(x). \tag{15}$$

It is easy to see that, since $\|r(\cdot)\|$ is continuous, $b_2>0$, $b_3>0$, and since $\epsilon(\cdot)$ is nonnegative and upper semicontinuous, then $\varphi(\cdot)$ is lower semicontinuous. We shall consider the level sets of $\varphi(\cdot)$ defined as

$$\mathscr{L}(\varphi, t) := \{x \in X \mid \varphi(x) \leq t\}, \qquad t \geq 0. \tag{16}$$

Note that the set $\mathscr{L}(\varphi, t)$ is closed for any $t \in \mathfrak{R}$ (Ref. 29, Theorem 7.1). Denoting $u = \|r(x)\|$, $\epsilon = \epsilon(x)$, and resolving the following quadratic inequality in u:

$$b_1 u^2 - b_2 \epsilon u - b_3 \epsilon^2 - t \leq 0,$$

we conclude that

$$u \leq b_2 \epsilon / 2b_1 + (1/2b_1)\sqrt{(b_2^2 + 4b_1 b_3)\epsilon^2 + 4b_1 t}.$$

Hence,

$$\mathscr{L}(\varphi, t) = \{x \in X \mid \|r(x)\| \leq b_2 \epsilon(x)/2b_1 + (1/2b_1)\sqrt{(b_2^2 + 4b_1 b_3)\epsilon(x)^2 + 4b_1 t}\}.$$

In particular,

$$\mathscr{L}(\varphi, 0) = \{x \in X \mid \|r(x)\| \leq [b_2 + \sqrt{b_2^2 + 4b_1 b_3}]\epsilon(x)/2b_1\}.$$

Defining

$$d_1 := [b_2 + \sqrt{b_2^2 + 4b_1 b_3}]/2b_1,$$
$$d_2 := b_1^{-1/2},$$

and taking into account the definition (6) of $S(\epsilon(\cdot))$, we further conclude that

$$\mathscr{L}(\varphi, t) \subset S(d_1 \epsilon(\cdot) + d_2 t^{1/2}), \tag{17}$$

$$\mathscr{L}(\varphi, 0) = S(d_1 \epsilon(\cdot)). \tag{18}$$

We next prove that there exists an accumulation point $\bar{x}$ of $\{x^i\}$ such that $\bar{x} \in \mathscr{L}(\varphi, 0)$. Suppose that the opposite holds. By (15), we have

$$f(x^i) - f(x^{i+1}) \geq \varphi(x^i), \qquad \forall i \geq i_1.$$

Since by our assumption

$$\overline{\mathrm{lt}}_{i \to \infty} \{x^i\} \cap \mathscr{L}(\varphi, 0) = \varnothing,$$

it follows from (16) and lower semicontinuity of $\varphi(\cdot)$ that, for some i_2 sufficiently large and some $c>0$,

$$\varphi(x^i) \geq c > 0, \qquad \forall i \geq i_2.$$

Denoting

$$k := \max\{i_1, i_2\}, \qquad \text{for any } i > k,$$

we have

$$f(x^k) - f(x^i) = \sum_{j=k}^{i-1} [f(x^j) - f(x^{j+1})] \geq \sum_{j=k}^{i-1} c = (i-k)c.$$

Letting $i \to \infty$, we get that $\{f(x^i)\} \to -\infty$, which contradicts the fact that $f(\cdot)$ is bounded from below on X. Hence, the assumption is invalid, and $lt_{i\to\infty}\{x^i\} \cap \mathscr{L}(\varphi, 0) \neq \varnothing$. Now, the first assertion of the theorem follows from (18).

Consider now a subsequence $\{x^{i_m}\}$ of $\{x^i\}$, and a $t \geq 0$ such that

$$\limsup_{m\to\infty} f(x^{i_m}) \leq \liminf_{i\to\infty} f(x^i) + t.$$

We shall establish that

$$\overline{\mathrm{lt}}_{m\to\infty} \{x^{i_m}\} \subset \mathscr{L}(\varphi, t).$$

Suppose this is not true. Then, passing onto a subsequence, if necessary, $\{x^{i_{m_k}}\} \to y \notin \mathscr{L}(\varphi, t)$. Therefore, by (16), for some $c > 0$,

$$\varphi(y) \geq t + 2c.$$

By the lower semicontinuity of $\varphi(\cdot)$, there exists k_1 sufficiently large such that

$$\varphi(x^{i_{m_k}}) \geq t + c, \qquad \forall k \geq k_1.$$

Let $k_2 := \min\{k \mid i_{m_k} \geq i_1\}$. By (15),

$$f(x^{i_{m_k}}) - f(x^{i_{m_k}+1}) \geq t + c, \qquad \forall k \geq \max\{k_1, k_2\}. \tag{19}$$

Also, since $\{x^{i_{m_k}}\} \to y$,

$$f(y) = \lim_{k\to\infty} f(x^{i_{m_k}}) \leq \limsup_{m\to\infty} f(x^{i_m}) \leq \liminf_{i\to\infty} f(x^i) + t. \tag{20}$$

Combining the last relation with (19), we have

$$\begin{aligned}\liminf_{i\to\infty} f(x^i) &\leq \liminf_{k\to\infty} f(x^{i_{m_k}+1})\\ &\leq \limsup_{k\to\infty} f(x^{i_{m_k}}) - t - c\\ &= \lim_{k\to\infty} f(x^{i_{m_k}}) - t - c\\ &= f(y) - t - c < f(y) - t,\end{aligned}$$

which contradicts (20). Hence,

$$\overline{\mathrm{lt}}_{m\to\infty}\ \{x^{i_m}\}\subset\mathscr{L}(\varphi, t),$$

and the second assertion of the theorem follows from (17).

For the last assertion, note that if the sequence $\{f(x^i)\}$ converges, then for every subsequence $\{x^{i_m}\}$ of $\{x^i\}$ it follows that

$$\limsup_{m\to\infty} f(x^{i_m})=\liminf_{i\to\infty} f(x^i).$$

Hence,

$$\overline{\mathrm{lt}}_{i\to\infty}\ \{x^i\}\subset\mathscr{L}(\varphi, 0),$$

and the last assertion of the theorem follows from (18). The proof is complete. □

Remark 2.1. If $\limsup_i \epsilon(x^i)\leq\epsilon$, and if the error bound (7) holds with $v\geq d_1\epsilon$, then it follows that there exist an accumulation point $\bar{x}$ of the sequence $\{x^i\}$ and a stationary point $\hat{x}\in S$ such that

$$\|\bar{x}-\hat{x}\|\leq\mu d_1\epsilon,$$

where μ is as specified in (7) and d_1 is given in Theorem 2.1.

3. Applications

In this section, we briefly discuss applications of our analysis to a number of well-known algorithms.

3.1. Gradient Projection Algorithm. We first consider the gradient projection algorithm (Refs. 2 and 3). In the presence of perturbations, it takes the following form:

$$x^{i+1}=[x^i-\eta_i\nabla f(x^i)+\delta(x^i, \eta_i)]^+.$$

Obviously, this method is a special case of Algorithm 2.1 corresponding to

$$e(x, \eta)=0, \qquad \forall x\in X.$$

Consequently, we can take $c_1=0$ and $c_2=0$ in (8)–(10). Provided the stepsize satisfies the standard conditions

$$0<c_3\leq\eta_i\leq 2/L-c_3,$$

it can be verified that

$$d_1 = O(L^2),$$

where d_1 is the constant involved in Theorem 2.1.

3.2. Proximal Minimization Algorithm. Given a current iterate x^i, the proximal minimization algorithm (Refs. 4 and 5) generates the next iterate x^{i+1} according to

$$x^{i+1} = \arg\min_{x \in X} \psi_i(x) := f(x) + (1/2\eta_i)\|x - x^i\|^2.$$

This method also falls within the presented framework as can be seen from the following. If the subproblems above are solved exactly, then the gradient projection optimality condition is satisfied, that is,

$$x^{i+1} = [x^{i+1} - c\nabla\psi_i(x^{i+1})]^+, \qquad \forall c > 0.$$

Suppose that only approximate solutions to the subproblems are available and that δ is the corresponding error. Then, we have

$$\begin{aligned} x^{i+1} &= [x^{i+1} - \eta_i \nabla\psi_i(x^{i+1}) + \delta(x^i, \eta_i)]^+ \\ &= [x^{i+1} - \eta_i\left(\nabla f(x^{i+1}) + \frac{1}{\eta_i}(x^{i+1} - x^i)\right) + \delta(x^i, \eta_i)]^+ \\ &= [x^i - \eta_i \nabla f(x^{i+1}) + \delta(x^i, \eta_i)]^+ \\ &= [x^i - \eta_i \nabla f(x^i) + e(x^i, \eta_i) + \delta(x^i, \eta_i)]^+, \end{aligned}$$

where

$$e(x^i, \eta_i) = \eta_i(\nabla f(x^i) - \nabla f(x^{i+1})).$$

By the Lipschitz continuity of the gradient,

$$\|e(x^i, \eta_i)\| \leq \eta_i L \|x^i - x^{i+1}\|;$$

hence, it is easy to see that (8)-(10) are satisfied provided

$$\limsup_i \eta_i < 1/L.$$

If $f(\cdot)$ is convex then

$$\langle e(x^i, \eta_i), x^i - x^{i+1}\rangle = \eta_i \langle \nabla f(x^i) - \nabla f(x^{i+1}), x^i - x^{i+1}\rangle \geq 0,$$

and we can further take $c_2 = 0$. It can be checked that

$$d_1 = O(L^2),$$

where d_1 is the constant involved in Theorem 2.1.

3.3. Extragradient Method. Consider now the extragradient method (Refs. 6 and 7), which updates a current iterate according to the double-projection formula

$$x^{i+1}=[x^i-\eta_i\nabla f([x^i-\eta_i\nabla f(x^i)]^+)]^+.$$

This iteration can be rewritten as

$$x^{i+1}=[x^i-\eta_i\nabla f(x^i)+e(x^i,\eta_i)]^+,$$

where

$$e(x^i,\eta_i)=\eta_i[\nabla f(x^i)-\nabla f([x^i-\eta_i\nabla f(x^i)]^+)].$$

In the presence of perturbations, we have

$$x^{i+1}=[x^i-\eta_i\nabla f(x^i)+e(x^i,\eta_i)+\delta(x^i,\eta_i)]^+,$$

where $\delta(x^i,\eta_i)$ is the aggregate perturbation at the ith iteration. Let

$$y^i=[x^i-\eta_i\nabla f(x^i)]^+.$$

By the Lipschitz continuity of the gradient, we have

$$\begin{aligned}\|e(x^i,\eta_i)\|&=\eta_i\|\nabla f(y^i)-\nabla f(x^i)\|\\&\le\eta_i L\|y^i-x^i\|.\end{aligned}$$

Furthermore,

$$\begin{aligned}\|x^{i+1}-x^i\|&\ge\|x^i-y^i\|-\|y^i-x^{i+1}\|\\&=\|x^i-y^i\|-\|[x^i-\eta_i\nabla f(x^i)]^+-[x^i-\eta_i\nabla f(y^i)]^+\|\\&\ge\|x^i-y^i\|-\eta_i\|\nabla f(x^i)-\nabla f(y^i)\|\\&\ge(1-\eta_i L)\|x^i-y^i\|,\end{aligned}$$

where the second inequality follows from Lemma 2.2 and the last inequality from the Lipschitz continuity of the gradient. Combining the last two relations, we obtain

$$\|e(x^i,\eta_i)\|\le[\eta_i L/(1-\eta_i L)]\|x^{i+1}-x^i\|.$$

It can be verified that conditions (8)–(10) are satisfied provided

$$\eta_i<1/2L.$$

3.4. Incremental Gradient Algorithms. Incremental algorithms are designed to solve the problem

$$\min_{x \in \Re^n} f(x) := \sum_{j=1}^{K} f_j(x)$$

of minimizing a finite summation of continuously differentiable (partial) objective functions f_j: $\Re^n \to \Re$, $j = 1, \ldots, K$, where the number K is typically large. This is an important problem in machine learning (in particular, neural network) applications, where weights and thresholds of the network comprise the problem variable $x \in \Re^n$, K is the number of training samples, and $f_j(\cdot)$ represents the error associated with the jth sample, $j = 1, \ldots, K$; see Ref. 30 for a detailed description.

In applications where K is large, the following incremental gradient algorithm proved to be very useful: having x^i, compute

$$x^{i+1} = T(x^i, \eta_i),$$

where T: $\Re^n \times \Re_+ \to \Re^n$ is given by

$$T(x, \eta) := x - \eta \sum_{j=1}^{K} \nabla f_j(z^j),$$

with

$$z^1 = x, \qquad z^{j+1} = z^j - \eta \nabla f_j(z^j), \quad j = 1, \ldots, K-1.$$

This algorithm processes partial objective functions one at a time and immediately updates the variables; hence, the connotation "incremental". On the domain of large neural network training problems, this algorithm is known to be often superior to standard optimization techniques which process all the partial objective functions before adjusting the variables. In particular, it is typically more effective than the standard gradient descent method, given by

$$\tilde{T}(x, \eta) := x - \eta \sum_{j=1}^{K} \nabla f_j(x) = x - \eta \nabla f(x).$$

Incremental methods constitute an area of active research; see Refs. 31–34, 17–18, and 20–21. We note that, in the above-mentioned papers, the stepsizes are chosen to satisfy the following condition:

$$\sum_{i=0}^{\infty} \eta_i = \infty, \qquad \sum_{i=0}^{\infty} \eta_i^2 < \infty.$$

This condition implies that the stepsizes tend to zero limit, while many heuristic rules used by practitioners keep them bounded away from zero.

The theory presented in this paper allows us to consider the computationally important case where

$$\liminf_{i} \eta_i \geq \bar{\eta} > 0.$$

We next show that the incremental gradient algorithm falls within our framework.

By the construction of the algorithm, we have

$$\begin{aligned} T(x, \eta) &= x - \eta \sum_{j=1}^{K} \nabla f_j(z^j) \\ &= x - \eta \left[\sum_{j=1}^{K} [\nabla f_j(z^j) - \nabla f_j(x) + \nabla f_j(x)] \right] \\ &= x - \eta \left[\sum_{j=1}^{K} \nabla f_j(x) + \sum_{j=1}^{K} [\nabla f_j(z^j) - \nabla f_j(x)] \right] \\ &= x - \eta \nabla f(x) + \delta(x, \eta), \end{aligned}$$

where

$$\delta(x, \eta) := -\eta \sum_{j=1}^{K} [\nabla f_j(z^j) - \nabla f_j(x)].$$

It is now clear that we have a special case of Algorithm 2.1. We refer the reader to Ref. 8 for a detailed analysis.

4. Concluding Remarks

A unified approach to the analysis of perturbed feasible descent methods has been presented. It was established that a certain ϵ-approximate solution can be obtained, where ϵ depends linearly on the level of perturbations. It is shown that the perturbed gradient projection, proximal minimization, extragradient and incremental gradient methods fall within the presented framework. Applications of the ideas presented here to other classes of optimization algorithms [for example, projection methods which are not descent methods [e.g., Refs. 35 and 36)] is an interesting subject of future research.

References

1. Luo, Z. Q., and Tseng, P., *Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach*, Annals of Operations Research, Vol. 46, pp. 157–178, 1993.

2. GOLDSTEIN, A. A., *Convex Programming in Hilbert Space*, Bulletin of the American Mathematical Society, Vol. 70, pp. 709-710, 1964.
3. LEVITIN, E. S., and POLYAK, B. T., *Constrained Minimization Methods*, USSR Computational Mathematics and Mathematical Physics, Vol. 6, pp. 1-50, 1965.
4. MARTINET, B., *Regularisation d'Inéquations Variationelles per Approximations Successives*, RAIRO-Operations Research, Vol. 4, pp. 154-159, 1970.
5. ROCKAFELLAR, R. T., *Monotone Operators and the Proximal Point Algorithm*, SIAM Journal on Control and Optimization, Vol. 14, pp. 877-898, 1976.
6. KORPELEVICH, G. M., *The Extragradient Method for Finding Saddle Points and Other Problems*, Matecon, Vol. 12, pp. 747-756, 1976.
7. MARCOTTE, P., *Application of Khobotov's Algorithm to Variational Inequalities and Network Equilibrium Problems*, Information Systems and Operational Research, Vol. 29, pp. 258-270, 1991.
8. SOLODOV, M. V., *Incremental Gradient Algorithms with Stepsizes Bounded Away from Zero*, Technical Report B-096, Instituto de Matematica Pura e Aplicada, Jardim Botanico, Rio de Janeiro, Brazil, 1995.
9. MANGASARIAN, O. L., *Convergence of Iterates of an Inexact Matrix Splitting Algorithm for the Symmetric Monotone Linear Complementarity Problem*, SIAM Journal on Optimization, Vol. 1, pp. 114-122, 1991.
10. LUO, Z. Q., and TSENG, P., *Error Bound and Convergence Analysis of Matrix Splitting Algorithms for the Affine Variational Inequality Problem*, SIAM Journal on Optimization, Vol. 2, pp. 43-54, 1992.
11. LI, W., *Remarks on Matrix Splitting Algorithms for Symmetric Linear Complementarity Problems*, SIAM Journal on Optimization, Vol. 3, pp. 155-163, 1993.
12. SOLODOV, M. V., *New Inexact Parallel Variable Distribution Algorithms*, Computational Optimization and Applications, (to appear).
13. DEMBO, R. S., EISENSTAT, S. C., and STEIHAUG, T., *Inexact Newton Methods*, SIAM Journal on Numerical Analysis, Vol. 19, pp. 400-408, 1982.
14. POLYAK, B. T., *Introduction to Optimization*. Optimization Software, Publications Division, New York, New York, 1987.
15. POLAK, E., *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, New York, 1971.
16. BOGGS, P. T., and DENNIS, J. E., *A Stability Analysis for Perturbed Nonlinear Iterative Methods*, Mathematics of Computation, Vol. 30, pp. 199-215, 1976.
17. 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.
18. 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.
19. 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.
20. SOLODOV, M. V., and ZAVRIEV, S. K., *Error-Stability Properties of Generalized Gradient-Type Algorithms*, Mathematical Programming Technical Report 94-05,

Computer Science Department, University of Wisconsin, Madison, Wisconsin, 1994 (Revised 1995).

21. 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.
22. MANGASARIAN, O. L., *Nonlinear Programming*, McGraw-Hill, New York, New York, 1969.
23. ROBINSON, S. M., *Some Continuity Properties of Polyhedral Multifunctions*, Mathematical Programming Study, Vol. 14, pp. 206–214, 1981.
24. LUO, Z. Q., MANGASARIAN, O. L., REN, J., and SOLODOV, M. V., *New Error Bounds for the Linear Complementarity Problem*, Mathematics of Operations Research, Vol. 19, pp. 880–892, 1994.
25. PANG, J. S., *A Posteriori Error Bounds for the Linearly-Constrained Variational Inequality Problem*, Mathematics of Operations Research, Vol. 12, pp. 474–484, 1987.
26. LUO, X. D., and TSENG, P., *On Global Projection-Type Error Bound for the Linear Complementarity Problem*, Linear Algebra and Its Applications, (to appear).
27. ZANGWILL, W. I., *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
28. GAFNI, E. M., and BERTSEKAS, D. P., *Two-Metric Projection Methods for Constrained Optimization*, SIAM Journal on Control and Optimization, Vol. 22, pp. 936-964, 1984.
29. ROCKAFELLAR, R. T., *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.
30. MANGASARIAN, O. L., *Mathematical Programming in Neural Networks*, ORSA Journal on Computing, Vol. 5, pp. 349–360, 1993.
31. BERTSEKAS, D. P., *Incremental Least Squares Methods and the Extended Kalman Filter*, SIAM Journal on Optimization, Vol. 6, pp. 807–822, 1996.
32. BERTSEKAS, D. P., *A New Class of Incremental Gradient Methods for Least Squares Problems*, Report, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1995.
33. 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.
34. TSENG, P., *Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule*, Report, Department of Mathematics, University of Washington, Seattle, Washington, 1995.
35. SOLODOV, M. V., and TSENG, P., *Modified Projection-Type Methods for Monotone Variational Inequalities*, SIAM Journal on Control and Optimization, Vol. 34, No. 5, 1996.
36. TSENG, P., *On Linear Convergence of Iterative Methods for the Variational Inequality Problem*, Journal of Computational and Applied Mathematics, Vol. 60, pp. 237–252, 1995.