SIAM J. OPTIM.
Vol. 8, No. 2, pp. 506–531, May 1998

012

AN INCREMENTAL GRADIENT(-PROJECTION) METHOD WITH MOMENTUM TERM AND ADAPTIVE STEPSIZE RULE*

PAUL TSENG†

Abstract. We consider an incremental gradient method with momentum term for minimizing the sum of continuously differentiable functions. This method uses a new adaptive stepsize rule that decreases the stepsize whenever sufficient progress is not made. We show that if the gradients of the functions are bounded and Lipschitz continuous over a certain level set, then every cluster point of the iterates generated by the method is a stationary point. In addition, if the gradient of the functions have a certain growth property, then the method is either linearly convergent in some sense or the stepsizes are bounded away from zero. The new stepsize rule is much in the spirit of heuristic learning rules used in practice for training neural networks via backpropagation. As such, the new stepsize rule may suggest improvements on existing learning rules. Finally, extension of the method and the convergence results to constrained minimization is discussed, as are some implementation issues and numerical experience.

Key words. incremental gradient method, gradient projection, convergence analysis, backpropagation, nonlinear neural network training

AMS subject classifications. 49M07, 49M37, 90C30

PII. S1052623495294797

1. Introduction.

Consider the problem of minimizing, over the n-dimensional real space $\Re^n$, a function $f : \Re^n \mapsto \Re$ of the form

$$f(x) = \sum_{i=1}^{m} f_i(x), \tag{1}$$

where f_i, $i = 1, ..., m$, are continuously differentiable functions from $\Re^n$ to $\Re$. Our interest in this problem stems from an important special case, that of nonlinear neural network training, in which x is the vector of weights in the neural network and $f_i(x)$ is the corresponding output error for the ith training example. (See [8], [9], [12], [13] for more detailed discussions of this connection.) Extension of our results to the constrained minimization of f will be discussed in section 5.

We will focus on the following iterative method for solving the preceding problem whereby, for a given $x_1^0 \in \Re^m$, we generate a sequence $\{(x_1^t, ..., x_{m+1}^t)\}_{t=0,1,...}$ according to

$$\begin{aligned} x_{i+1}^t &:= x_i^t - \alpha^t d_i^t, \quad i = 1, ..., m, \\ x_1^{t+1} &:= x_{m+1}^t, \end{aligned} \tag{2}$$

where α^t is a positive scalar (the "stepsize") and

$$d_i^t := \begin{cases} \nabla f_i(x_i^t) + \zeta d_m^{t-1} & \text{if } i = 1, \\ \nabla f_i(x_i^t) + \zeta d_{i-1}^t & \text{if } i > 1 \end{cases} \tag{3}$$

with $d_m^{-1} = 0$ and $\zeta \in [0, 1)$. Here, each direction d_i^t is a weighted sum of the previous direction (the "momentum term") and the gradient of f_i at x_i^t. Thus, unlike

*Received by the editors November 15, 1995; accepted for publication (in revised form) January 30, 1997. This work was supported by National Science Foundation grant CCR-9311621.
http://www.siam.org/journals/siopt/8-2/29479.html
†Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu).

conventional gradient methods, this method does not use the gradient of f to take a step but only the gradient of one of the f_i. In the special case where $m = 1$, this method reduces to the steepest descent method for $\zeta = 0$ and to the heavy-ball method [17, p. 65] for $\zeta \geq 0$. For general m and $\zeta = 0$, this method is reminiscent of a nonlinear least square algorithm of Davidon [5], which has been further studied by Pappas [15] and Bertsekas [2]. When applied to neural network training, this method reduces to the very popular on-line backpropagation algorithm with a momentum term (with "training/learning rates" identified with stepsizes), as conceived by Werbos [20], Le Cun [11], Parker [16], and Rumelhart, Hinton, and Williams [18] (see the discussions in [21]). Numerical experience suggests that it is typically beneficial to choose $\zeta > 0$. (In [18, p. 330], a value of $\zeta \approx .9$ is recommended.) An interesting one-parameter generalization of this method for $\zeta = 0$ and of the steepest descent method is recently studied in [3].

A key issue concerns the stepsizes $\{\alpha^t\}_{t=0,1,...}$ which, to quote from [14], "are often crucial for the success of the algorithm." In the case of neural network training, various heuristic rules for choosing the stepsize have been proposed, the most popular of which entail keeping the stepsize fixed for as long as "progress" is made and decreasing the stepsize if otherwise. However, these heuristic rules are justified only by extensive experimentation (see [10, p. 124], [19], and references therein and in [6]). More recently, stepsize rules have been proposed for the special case of $\zeta = 0$, for which global convergence can be shown under mild assumptions on $f, f_1, ..., f_m$. One such rule, studied in [3], [8], [12], [13], [22], [23], requires the stepsizes $\{\alpha^t\}_{t=0,1,...}$ to be square summable but not summable, i.e.,

$$\sum_{t=0}^{\infty} (\alpha^t)^2 < \infty, \qquad \sum_{t=0}^{\infty} \alpha^t = \infty.$$

The reference [8] also considers the more general stepsize rule in which square summability of $\{\alpha^t\}_{t=0,1,...}$ is replaced by $\alpha^t \to 0$. These rules, however, always require the stepsize to tend to zero. To see the drawback of this, suppose $f_1, ..., f_m$ are identical with Lipschitz continuous gradients. Then the method (2)–(3) with $\zeta = 0$ is just the steepest descent method for minimizing f, for which it is well known that convergence does not require the stepsize to tend to zero (and that a stepsize tending to zero yields slow convergence). A second rule, proposed in [12], requires that

$$\alpha^t \propto \left\| \sum_{i=1}^{m} \nabla f_i(x_i^{t-1}) \right\|^2 .$$

This rule uses information about f, but it still requires the stepsize to tend to zero (the right-hand side tends to zero as $x_1^{t-1}, ..., x_m^{t-1}$ all tend to a stationary point of f) and in practice the convergence seems to be slow. A third rule, proposed in [9], chooses α^t to be the largest element of $\{\alpha^{t-1}, \omega\alpha^{t-1}, \omega^2\alpha^{t-1}, ...\}$ for which the following sufficient descent condition is satisfied:

$$f(x_1^{t+1}) \leq f(x_1^t) - \epsilon_1 \alpha^t \left\| \sum_{i=1}^{m} \nabla f_i(x_i^t) \right\|^2 - \epsilon_2 (\alpha^t)^2 \sum_{i=1}^{m} \left\| \nabla f_i(x_i^t) \right\|^2 ,$$

where $\omega \in (0, 1)$, $\epsilon_1 \in (0, 1)$, $\epsilon_2 > 0$ are parameters. This rule uses information about f and does not always require the stepsize to tend to zero. Moreover, it is in the spirit of heuristic rules used in practice in that it keeps the stepsize fixed for as long

as sufficient descent is made and decreases the stepsize if otherwise. On the other hand, this rule still tends to make the stepsize very small since it requires sufficient descent at every iteration and when the term $\sum_{i=1}^m \|\nabla f_i(x_i^t)\|^2$ is bounded away from zero, the stepsize must necessarily tend to zero. In addition, the preceding stepsize rules apply to the case of $\zeta = 0$. It is unclear whether these rules can be extended to the case of $\zeta > 0$, which is the case of practical interest. (The work of [13] considers a version of the method that uses a momentum term. However, the momentum term uses only the history of the method from the start of the current iteration.)

In this paper, we propose a new rule (see (5)–(6) and (7)–(9)) for choosing the stepsizes $\{\alpha^t\}_{t=0,1,...}$ for which convergence of the method (2)–(3) can be shown for any $\zeta \in [0, (.5)^{1/m})$. (Note that $(.5)^{1/m} > .9$ for $m \geq 8$, so the restriction on ζ is mild.) This new rule, like the rule proposed in [9] for the case of $\zeta = 0$, does not always require the stepsize to tend to zero and keeps the stepsize fixed for as long as descent in an overall sense is achieved and decreases the stepsize if otherwise. Unlike the rule of [9], this new stepsize rule does not require descent at every iteration and, as such, the stepsize tends to remain large which is essential for good convergence.

We show that the method (2)–(3) using this stepsize rule has desirable global convergence properties (see Proposition 3.4). Moreover, in the case where $\nabla f_1, ..., \nabla f_m$ grow at most linearly in norm with ∇f (which, in the context of neural network training, amounts to the neural network being trainable so to achieve zero output error on the training examples), either the method is linearly convergent in some sense or the stepsize is bounded away from zero (see Proposition 4.2). The method and the convergence results can also be extended, with suitable modifications, to the problem of constrained minimization of f (see section 5). We note that neither $d_1^t, ..., d_m^t$ nor their sum need be a descent direction for f, so conventional convergence arguments cannot be applied here. Moreover, for $\zeta > 0$, the proofs are further complicated by the dependence of d_i^t on the entire past history of the method up to then. And while some of our proof ideas are adapted from [12] and [13], much of the arguments are new due to the use of a new stepsize rule and the presence of the momentum term. In section 6, we discuss implementation issues and numerical experience with the method.

A few words about our notations: For any $x, y \in \Re^n$, $\langle x, y\rangle$ denotes the usual inner product of vectors x and y and $\|x\|$ denotes the Euclidean norm of x, i.e., $\|x\| = \langle x, x\rangle^{1/2}$. For any continuously differentiable function $h : \Re^n \mapsto \Re$, we say that ∇h, the gradient of h, is Lipschitz continous (with constant $\lambda \geq 0$) on a subset X of $\Re^n$ if

$$\|\nabla h(x) - \nabla h(x')\| \ \leq \ \lambda \|x - x'\| \quad \forall x, \ x' \in X.$$

For any two positive integers $t > \tau$ and any scalar $\theta > 0$, we denote

$$\theta_\tau^t = \theta^{\tau m} + \ \theta^{\tau m+1} + \cdots + \theta^{tm-1}.$$

Thus, $\theta_s^\tau + \theta_\tau^t = \theta_s^t$ for any $t > \tau > s$. For any $i \in \{1, ..., m\}$ and positive integer j, we define

$$i \ominus j := (i - j - 1 \bmod m - 1) + 1.$$

Thus $i \ominus j = i - j$ for $j = 1, ..., i - 1$ and $i \ominus j = i - j + m$ for $j = i, ..., i + m - 1$ and so on.

2. Method description. In this section, we describe in detail the method (2)–(3) for the unconstrained minimization of f given by (1) and the new rule for choosing the stepsize α^t adaptively. To describe this new stepsize rule, we follow [12] and make the following assumption about $f, f_1, ..., f_m$ and the initial iterate x_1^0.

Assumption A. There exist scalars $\eta > f(x_1^0)$ and $\rho > 0$ such that, for $i = 1, ..., m$, ∇f_i is bounded and Lipschitz continuous (with some constant $\lambda_i \geq 0$) on the set

$$\mathcal{R}_\rho^\eta := \{\ x \in \Re^n\ :\ f(x) \leq \eta\ \} + \rho\mathcal{B},$$

where $\mathcal{B} = \{\ x \in \Re^n\ :\ \|x\| \leq 1\ \}$.

Assumption A is quite mild and, in particular, is satisfied when $f_1, ..., f_m$ are twice differentiable and the level set $\{\ x \in \Re^n : f(x) \leq \eta\ \}$ is bounded for some $\eta > f(x_1^0)$, as is the typical case with neural network training. (See [12, section 3] for further discussions.)

The new stepsize rule for choosing α^t depends on η, ρ, and $\lambda_1, ..., \lambda_m$ and, in the spirit of the Armijo–Goldstein stepsize rule for gradient descent methods, periodically checks if a certain descent condition is satisfied since the previous check was made and, if not, decreases the stepsize and restarts the method from when the previous check was made. Below, we formally state the method (2)–(3) using this stepsize rule. Following [3], we will call this method the *incremental gradient method*.

Incremental gradient method (with momentum term). Choose any $x_1^0 \in \Re^n$ such that Assumption A holds for some η, ρ and $\lambda_1, ..., \lambda_m$. Choose any $\zeta \in [0, (.5)^{1/m})$ and let

$$\delta_1 := \frac{1 - 2\zeta^m}{1 - \zeta}, \qquad \delta_2 := \frac{.5\zeta^m}{1 - \zeta}, \qquad \delta_3 := .5(\lambda_1 + \cdots + \lambda_m)(1 + \zeta^m). \tag{4}$$

(By choice of ζ, we have $\delta_1 > 0$.) Choose any $\omega \in (0, 1)$ and any subsequence T of $\{1, 2, ...\}$ containing 1. Choose any positive scalars $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3$ satisfying $\epsilon_1 < \delta_1$ and $\epsilon_2 \frac{\zeta^m}{1-\zeta^m} < \eta - f(x_1^0)$.

Step 0. Let α^0 be the largest element of $\{\epsilon_0, \omega\epsilon_0, \omega^2\epsilon_0, ...\}$, for which $x_2^0, ..., x_m^0, x_1^1$ given by (2)–(3) with $t = 0$ satisfy the following two conditions:

$$(x_2^0, ..., x_m^0, x_1^1) \in (\mathcal{R}_\rho^\eta)^m, \tag{5}$$

$$f(x_1^1) \leq \eta - \delta_3(\alpha^0\beta^0)^2 - \epsilon_2\frac{\zeta^m}{1 - \zeta^m}, \tag{6}$$

where β^0 is given by (10).

Step 1. For each $s \in T$, let $h = \min\{t \in T : t > s\}$ and we generate $(x_1^h, d_m^{h-1}, \alpha^{h-1})$ from $(x_1^s, d_m^{s-1}, \alpha^{s-1})$ as follows: If $\nabla f(x_1^s) = 0$, we stop. Else, we let α^{h-1} be the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, \omega^2\alpha^{s-1}, ...\}$ for which $x_2^t, ..., x_m^t, x_1^{t+1}$ given by (2)–(3) with $\alpha^t = \alpha^{h-1}$, $t = s, ..., h-1$, satisfy the following three conditions:

$$(x_2^t, ..., x_m^t, x_1^{t+1}) \in (\mathcal{R}_\rho^\eta)^m, \quad t = s, ..., h - 1, \tag{7}$$

$$f(x_1^h) \leq \eta - (\delta_2 + \epsilon_1)\, p^h - \delta_3 q^h + \delta_2(1 - \zeta)u^h + \delta_3(1 - \zeta)v^h - \epsilon_2\frac{\zeta^{hm}}{1 - \zeta^m}, \tag{8}$$

$$\|\nabla f(x_1^s) - g^s\| \leq \epsilon_3\|g^s\|, \tag{9}$$

where, for $t = s, ..., h - 1$, we define

$$g^t := \sum_{i=1}^m \nabla f_i(x_i^t), \qquad \beta^t := \max\left\{\|g^t\|, \sum_{i=1}^m \|d_i^t\|\right\}, \tag{10}$$

(11)

$$p^{t+1} := p^t + \alpha^t\|g^t\|^2, \quad u^{t+1} := u^t + a^t, \quad a^{t+1} := \zeta^m a^t + (1+\cdots+\zeta^{m-1})\alpha^t\|g^t\|^2,$$
$$q^{t+1} := q^t + (\alpha^t\beta^t)^2, \quad v^{t+1} := v^t + b^t, \quad b^{t+1} := \zeta^m b^t + (1+\cdots+\zeta^{m-1})(\alpha^t\beta^t)^2,$$

with $p^1 := u^1 := a^1 := q^0 := v^0 := b^0 := 0$.

Roughly speaking, the stepsize rule checks at each iteration $s \in T$ whether the current stepsize is acceptable (i.e., satisfies (7)–(9)) for all iterations between s and the next element h of T and, if not, it decreases the stepsize by ω and restarts the method from iteration s. Thus, if the elements of T are far apart, then the rule makes this check infrequently but it needs to backtrack further whenever the stepsize needs to be decreased. In our testing (see section 6), checking every 10 iterations, i.e., $T = \{1, 11, 21, 31, ...\}$, worked well. A more sophisticated strategy might be to check more frequently at the beginning, e.g., $T = \{1, 2, 5, 11, 21, 31, ...\}$. We remark that we can also increase the stepsize, provided that this is done only a finite number of times. For the other parameters in the stepsize rule, it suffices to choose ϵ_1, ϵ_2 reasonably small, choose ϵ_3, η, ρ reasonably large, choose ϵ_0 near 1, and choose ω, ζ away from 0 and 1. Only the parameter $\lambda_1 + \cdots + \lambda_m$, which is problem dependent, requires significant fine tuning (if this is too large, then the stepsize becomes too small; if this is too small, then the stepsize remains large but the method experiences large oscillations). In our testing (see section 6), the choice of $\epsilon_1 = \epsilon_2 = .00001$, $\epsilon_3 = 1000, \eta = 1.5f(x_1^0) + 100, \rho = \infty$, $\epsilon_0 = 1$, $\omega = .5$, $\zeta = .8$, and $\lambda_1 + \cdots + \lambda_m = 1$ worked well. However, to solve a wider range of problems, we would need to update λ_i on-line by, for example, $\|\nabla f_i(x_i^1) - \nabla f_i(x_i^0)\|/\|x_i^1 - x_i^0\|$, for $i = 1, ..., m$.

Of the three conditions (7)–(9), both (7) and (9) are quite unrestrictive since ρ is typically large (e.g., in [8] and [13], it is assumed that $\rho = \infty$) and we can choose ϵ_3 arbitrarily large. To see that (8) is also unrestrictive, first note that (8) does not require f-value to be monotonically decreasing in any sense but only that the f-value be less than η by some positive quantity. This positive quantity, which is easily computable using the updating formula (11), depends on $f, f_1, ..., f_m$ and is increasing with t only in a long-run sense. More precisely, by using a straightforward calculation from (11), we see that

(12)

$$p^t = \sum_{\tau=1}^{t-1} \alpha^\tau\|g^\tau\|^2, q^t = \sum_{\tau=0}^{t-1} (\alpha^\tau\beta^\tau)^2, u^t = \sum_{\tau=1}^{t-2} \zeta_0^{t-\tau-1}\alpha^\tau\|g^\tau\|^2, v^t = \sum_{\tau=0}^{t-2} \zeta_0^{t-\tau-1}(\alpha^\tau\beta^\tau)^2.$$

Since $\zeta_0^{t-\tau-1} \approx 1/(1-\zeta)$ for $\tau = 0, ..., t-2$ (this is true especially for large m), then $p^t \approx (1-\zeta)u^{t+1}$ and $q^t \approx (1-\zeta)v^{t+1}$, so the right-hand side of (8) may increase or decrease with h, depending on how $\alpha^\tau\|g^\tau\|^2$ and $(\alpha^\tau\beta^\tau)^2$ change with τ, though in the long run the tendency is towards a decrease. To see this, we note from (12) and $\zeta_0^{t-\tau-1} \le 1/(1-\zeta)$ for $\tau = 0, ..., t-2$ that (8) implies

$$f(x_1^h) \le \eta - \epsilon_1 \sum_{\tau=1}^{h-1} \alpha^\tau\|g^\tau\|^2 - \delta_2\alpha^{h-1}\|g^{h-1}\|^2 - \delta_3(\alpha^{h-1}\beta^{h-1})^2 - \epsilon_2\frac{\zeta^{hm}}{1-\zeta^m}. \tag{13}$$

The second term on the right-hand side, which is the dominant term there, is decreasing with h.

3. Global convergence analysis. In this section we show that the incremental gradient method of section 2 has desirable global convergence properties (see Proposition 3.4). Throughout, we will assume that Assumption A holds. First, we have the following technical lemma.

LEMMA 3.1. *For any $t \in \{0, 1, ...\}$, any $\alpha^t > 0$ and any $x_1^t, ..., x_{m+1}^t$ in $\mathcal{R}_\rho^\eta$ satisfying* (2)*, we have*

$$\|\nabla f_i(x_j^t) - \nabla f_i(x_i^t)\| \le \lambda_i \alpha^t \beta^t,$$

for $1 \le i, j \le m+1$, where β^t is given by (10)*.*

Proof. By (2) and (10), for any $1 \le j \le i \le m+1$, we have

$$\|x_j^t - x_i^t\| = \alpha^t \left\| \sum_{l=j}^{i-1} d_l^t \right\| \le \alpha^t \beta^t.$$

2A similar argument shows the above inequality also holds for any $1 \le i \le j \le m+1$. Since $x_1^t, ..., x_{m+1}^t$ are in $\mathcal{R}_\rho^\eta$, the above inequality, together with ∇f_i being Lipschitz continuous (with constant λ_i) on $\mathcal{R}_\rho^\eta$ for $i = 1, ..., m$, yields the desired inequality. □

Under Assumption A, there exist positive scalars $\beta_1, ..., \beta_m$ such that

$$\|\nabla f_i(x)\| \le \beta_i \qquad \forall x \in \mathcal{R}_\rho^\eta, \qquad i = 1, ..., m. \tag{14}$$

Let

$$\beta \;:=\; \beta_1 + \cdots + \beta_m, \qquad \lambda \;:=\; \lambda_1 + \cdots + \lambda_m. \tag{15}$$

The next lemma shows that, for α^t sufficiently small, $x_1^t, ..., x_{m+1}^t$ satisfying (2)–(3) remain in $\mathcal{R}_\rho^\eta$ for $t = 0, 1, ...$

LEMMA 3.2. *For any $t \in \{0, 1, ...\}$, any $\alpha^\tau \in (0, \rho(1-\zeta)/\beta]$ and any $x_1^\tau, ..., x_{m+1}^\tau$ satisfying* (2)–(3) *for $\tau = 0, 1, ..., t$, and such that $(x_1^\tau, ..., x_m^\tau) \in (\mathcal{R}_\rho^\eta)^m$ for $\tau = 0, 1, ..., t-1$ and $f(x_1^t) \le \eta$, we have that $x_1^t, ..., x_m^t$ are in $\mathcal{R}_\rho^\eta$, as is the line segment joining x_1^t with x_1^{t+1}.*

Proof. First, we claim that

$$\|x_l^t - x_1^t\| \le \frac{\rho(1-\zeta)}{\beta} \sum_{j=1}^{l-1} \sum_{k=0}^{tm+j-1} \zeta^k \beta_{j \ominus k}, \tag{16}$$

for $l = 1, ..., m+1$. We prove this by induction on l. Clearly, (16) holds for $l = 1$. Suppose (16) holds for $l = 1, ..., i$ for some $i \in \{1, ..., m\}$ and we show below that it also holds for $l = i+1$. For $l = 1, ..., i$, since (16) holds and the right-hand side of (16) is bounded above by ρ, we have $\|x_l^t - x_1^t\| \le \rho$ so that (cf. $f(x_1^t) \le \eta$) $x_l^t \in \mathcal{R}_\rho^\eta$. Then, (2)–(3) yields

$$\begin{aligned} \|x_{i+1}^t - x_i^t\| = \alpha^t \left\| \sum_{k=0}^{tm+i-1} \zeta^k \nabla f_{i \ominus k} \left(x_{i \ominus k}^{\lfloor \frac{tm+i-1-k}{m} \rfloor} \right) \right\| &\le \alpha^t \sum_{k=0}^{tm+i-1} \zeta^k \beta_{i \ominus k} \\ &\le \frac{\rho(1-\zeta)}{\beta} \sum_{k=0}^{tm+i-1} \zeta^k \beta_{i \ominus k}, \end{aligned}$$

where the first inequality follows from $x_1^0, x_2^0, ..., x_{i-1}^t, x_i^t \in \mathcal{R}_\rho^\eta$ and (14). Since (16) holds for $l = i$, this shows that (16) holds for $l = i + 1$.

Since (16) holds and the right-hand side of (16) is bounded above by ρ, $f(x_1^t) \le \eta$ implies $x_l^t \in \mathcal{R}_\rho^\eta$ for $l = 1, ..., m+1$. Moreover, (16) with $l = m+1$ and (2) implies $\|x_1^{t+1} - x_1^t\| \le \rho$, so the line segment joining x_1^t with x_1^{t+1} lies in $\mathcal{R}_\rho^\eta$. □

By using Lemma 3.1, we obtain the following lemma estimating the decrease in f-value per iteration of the incremental gradient method.

LEMMA 3.3. *For any $t \in \{1, 2, ...\}$, any $\alpha^\tau > 0$ and any $x_1^\tau, ..., x_{m+1}^\tau$ in $\mathcal{R}_\rho^\eta$ satisfying (2)–(3) for $\tau = 0, 1, ..., t$, and such that the line segment joining x_1^t with x_1^{t+1} lies in $\mathcal{R}_\rho^\eta$, we have*

$$f(x_1^{t+1}) \le f(x_1^t) - (\delta_1 + \delta_2)\alpha^t\|g^t\|^2 + \lambda(1.5 + 2\zeta\zeta_0^\infty)(\alpha^t\beta^t)^2$$

$$+ \delta_2(1-\zeta)\sum_{\tau=1}^{t-1}\zeta_{t-\tau-1}^{t-\tau}\alpha^t\|g^\tau\|^2 + \delta_3(1-\zeta)\sum_{\tau=0}^{t-1}\zeta_{t-\tau-1}^{t-\tau}(\alpha^\tau\beta^\tau)^2 + \alpha^t\|g^t\|\zeta^{tm}\bar{\beta}, \tag{17}$$

where $\delta_1, \delta_2, \delta_3$ are given by (4), g^t, β^t are given by (10), and $\bar{\beta} := \|\sum_{j=1}^m \sum_{i=j}^m \zeta^{i-j}\nabla f_j(x_j^0)\|$. Similarly, for any $\alpha^0 > 0$ and any $x_1^0, ..., x_m^0, x_1^1$ in $\mathcal{R}_\rho^\eta$ satisfying (2)–(3) (with $t = 0$) and such that the line segment joining x_1^0 with x_1^1 lies in $\mathcal{R}_\rho^\eta$, we have

$$f(x_1^1) \le f(x_1^0) - \alpha^0\|g^0\|^2 + 1.5\lambda(\alpha^0\beta^0)^2 + 2\alpha^0\|g^0\|\beta^0. \tag{18}$$

Proof. Fix any $t \in \{1, 2, ...\}$. Since the line segment joining x_1^t with x_1^{t+1} lies in $\mathcal{R}_\rho^\eta$ and, by $\nabla f = \nabla f_1 + \cdots + \nabla f_m$ (cf. (1)) and Assumption A, ∇f is Lipschitz continuous (with constant λ given by (15)) on $\mathcal{R}_\rho^\eta$, we obtain from the intermediate value theorem (see [4, p. 639]) that

$$f(x_1^{t+1}) \;\le\; f(x_1^t) + \langle \nabla f(x_1^t), x_1^{t+1} - x_1^t\rangle + ..5\lambda\|x_1^{t+1} - x_1^t\|^2. \tag{19}$$

Using (2) and (10), we bound the rightmost term in (19) as follows:

$$\|x_1^{t+1} - x_1^t\| = \alpha^t\left\|\sum_{i=1}^m d_i^t\right\| \le \alpha^t\beta^t. \tag{20}$$

The second term on the right-hand side of (19) can be bounded as follows:

$$\begin{aligned}
&\langle \nabla f(x_1^t), x_1^{t+1} - x_1^t\rangle \\
&= \langle \nabla f(x_1^t) - g^t, x_1^{t+1} - x_1^t\rangle + \left\langle g^t, -\alpha^t\sum_{i=1}^m d_i^t + \alpha^t\zeta_0^1 g^t\right\rangle - \alpha^t\zeta_0^1\|g^t\|^2 \\
&\le \|\nabla f(x_1^t) - g^t\|\|x_1^{t+1} - x_1^t\| + \alpha^t\|g^t\|\left\|\sum_{i=1}^m d_i^t - \zeta_0^1 g^t\right\| - \alpha^t\zeta_0^1\|g^t\|^2, \qquad (21)
\end{aligned}$$

where the equality follows from (2). Also, we have from (3) and (10) that

$$
\begin{aligned}
&\left\|\sum_{i=1}^{m} d_i^t - \zeta_0^1 g^t\right\| \\
&= \left\|\sum_{i=1}^{m}\left(\sum_{\tau=0}^{t-1}\left(\sum_{j=m-i+1}^{m}\zeta^{(t-\tau)m-j}\nabla f_{i+j-m}(x_{i+j-m}^{\tau+1})\right.\right.\right. \\
&\left.\left.\left. + \sum_{j=1}^{m-i}\zeta^{(t-\tau)m-j}\nabla f_{i+j}(x_{i+j}^{\tau})\right) + \sum_{j=1}^{i}\zeta^{tm+i-j}\nabla f_j(x_j^0)\right) - \zeta_0^1 g^t\right\| \\
&= \left\|\sum_{\tau=0}^{t-1}\left(\sum_{j=1}^{m}\zeta^{(t-\tau)m-j}\left(\sum_{k=1}^{j}\nabla f_k(x_k^{\tau+1}) + \sum_{k=j+1}^{m}\nabla f_k(x_k^{\tau})\right)\right)\right. \\
&\left. + \sum_{j=1}^{m}\sum_{i=j}^{m}\zeta^{tm+i-j}\nabla f_j(x_j^0) - \zeta_0^1 g^t\right\| \\
&= \left\|\sum_{\tau=0}^{t-1}\sum_{j=1}^{m}\zeta^{(t-\tau)m-j} g_j^{\tau+1} + \sum_{j=1}^{m}\sum_{i=j}^{m}\zeta^{tm+i-j}\nabla f_j(x_j^0) - \sum_{j=1}^{m}\zeta^{m-j} g^t\right\| \\
&= \left\|\sum_{\tau=0}^{t-2}\sum_{j=1}^{m}\zeta^{(t-\tau)m-j} g_j^{\tau+1} + \sum_{j=1}^{m}\sum_{i=j}^{m}\zeta^{tm+i-j}\nabla f_j(x_j^0) + \sum_{j=1}^{m}\zeta^{m-j}\left(g_j^t - g^t\right)\right\| \\
&\le \sum_{\tau=0}^{t-2}\sum_{j=1}^{m}\zeta^{(t-\tau)m-j}(\|g^{\tau+1}\| + \|g_j^{\tau+1} - g^{\tau+1}\|) + \left\|\sum_{j=1}^{m}\sum_{i=j}^{m}\zeta^{tm+i-j}\nabla f_j(x_j^0)\right\| \\
&\quad + \sum_{j=1}^{m}\zeta^{m-j}\|g_j^t - g^t\| \\
&= \sum_{\tau=0}^{t-2}\sum_{j=1}^{m}\zeta^{(t-\tau)m-j}\|g^{\tau+1}\| + \sum_{\tau=0}^{t-1}\sum_{j=1}^{m-1}\zeta^{(t-\tau)m-j}\|g_j^{\tau+1} - g^{\tau+1}\| + \zeta^{tm}\bar{\beta} \\
&\le \sum_{\tau=0}^{t-2}\sum_{j=1}^{m}\zeta^{(t-\tau)m-j}\|g^{\tau+1}\| + \lambda\sum_{\tau=0}^{t-1}\sum_{j=0}^{m-1}\zeta^{(t-\tau)m-j}(\alpha^{\tau+1}\beta^{\tau+1} + \alpha^{\tau}\beta^{\tau}) + \zeta^{tm}\bar{\beta} \\
&= \sum_{\tau=1}^{t-1}\zeta_{t-\tau}^{t-\tau+1}\|g^{\tau}\| + \lambda\zeta\left(\sum_{\tau=0}^{t-1}\zeta_{t-\tau-1}^{t-\tau+1}\alpha^{\tau}\beta^{\tau} + \zeta_0^1\alpha^t\beta^t - \zeta_{t-1}^t\alpha^0\beta^0\right) + \zeta^{tm}\bar{\beta}, \qquad (22)
\end{aligned}
$$

where the second equality follows from interchanging the order of the summations over i and j and then making the substitution $k = i + j - m$ (respectively, $k = i + j$) in the first (respectively, second) summation inside the doubly nested parentheses; the third equality follows by letting

$$
g_j^{\tau+1} := \sum_{k=1}^{j}\nabla f_k(x_k^{\tau+1}) + \sum_{k=j+1}^{m}\nabla f_k(x_k^{\tau}); \qquad (23)
$$

the fifth equality uses the fact $g_m^{\tau+1} = g^{\tau+1}$; the second inequality follows from (15) and the following consequence of (10) and Lemma 3.1 (since $x_1^\tau, ..., x_{m+1}^\tau = x_1^{\tau+1}, ..., x_m^{\tau+1} \in \mathcal{R}_\rho^\eta$):

$$\|g_j^{\tau+1} - g^{\tau+1}\| = \left\| \sum_{k=j+1}^{m} \left(\nabla f_k(x_k^\tau) - \nabla f_k(x_{m+1}^\tau) + \nabla f_k(x_1^{\tau+1}) - \nabla f_k(x_k^{\tau+1}) \right) \right\|$$
$$\leq \sum_{k=j+1}^{m} \lambda_k (\alpha^\tau \beta^\tau + \alpha^{\tau+1} \beta^{\tau+1})$$

for $j = 1, ..., m-1$ and for $\tau = 0, 1, ..., t-1$. By a similar argument, we have that

$$\|\nabla f(x_1^t) - g^t\| = \left\| \sum_{i=1}^{m} (\nabla f_i(x_1^t) - \nabla f_i(x_i^t)) \right\| \leq \sum_{i=1}^{m} \lambda_i \alpha^t \beta^t = \lambda \alpha^t \beta^t. \tag{24}$$

Using (20)–(24) to bound the right-hand side of (19) and then using β^t to bound $\|g^t\|$ (cf. (10)) yields

$$f(x_1^{t+1}) \leq f(x_1^t) - \zeta_0^1 \alpha^t \|g^t\|^2 + 1.5\lambda(\alpha^t \beta^t)^2 + \alpha^t \|g^t\| \sum_{\tau=1}^{t-1} \zeta_{t-\tau}^{t-\tau+1} \|g^\tau\|$$
$$+ \alpha^t \beta^t \lambda \zeta \left(\sum_{\tau=0}^{t-1} \zeta_{t-\tau-1}^{t-\tau+1} \alpha^\tau \beta^\tau + \zeta_0^1 \alpha^t \beta^t \right) + \alpha^t \|g^t\| \zeta^{tm} \bar{\beta}$$
$$\leq f(x_1^t) - \zeta_0^1 \alpha^t \|g^t\|^2 + 1.5\lambda(\alpha^t \beta^t)^2 + .5\alpha^t \left(\zeta_1^t \|g^t\|^2 + \sum_{\tau=1}^{t-1} \zeta_{t-\tau}^{t-\tau+1} \|g^\tau\|^2 \right)$$
$$+ \lambda \zeta \left(..5 \left((\zeta_0^t + \zeta_1^{t+1})(\alpha^t \beta^t)^2 + \sum_{\tau=0}^{t-1} \zeta_{t-\tau-1}^{t-\tau+1} (\alpha^\tau \beta^\tau)^2 \right) \right.$$
$$\left. + \zeta_0^1 (\alpha^t \beta^t)^2 \right) + \alpha^t \|g^t\| \zeta^{tm} \bar{\beta}$$
$$= f(x_1^t) - \left(\zeta_0^1 - .5\zeta_1^t \right) \alpha^t \|g^t\|^2 + \lambda(1.5 + .5\zeta\zeta_0^t + .5\zeta\zeta_1^{t+1} + \zeta\zeta_0^1)(\alpha^t \beta^t)^2$$
$$+ .5\alpha^t \zeta^m \sum_{\tau=1}^{t-1} \zeta_{t-\tau-1}^{t-\tau} \|g^\tau\|^2 + .5\lambda\zeta(1 + \zeta^m) \sum_{\tau=0}^{t-1} \zeta_{t-\tau-1}^{t-\tau} (\alpha^\tau \beta^\tau)^2 + \alpha^t \|g^t\| \zeta^{tm} \bar{\beta},$$

where the second inequality follows from using $ab \leq .5(a^2 + b^2)$ and the equality uses properties of ζ_i^j. Bounding from above ζ_1^t by $\zeta_1^\infty = \zeta^m/(1-\zeta)$ and $\zeta_0^t, \zeta_1^{t+1}, \zeta_0^1$ by ζ_0^∞ in the above expression and then using (4) yields (17). The proof of (18) is very similar. □

Below, we state and prove the main global convergence result for the incremental gradient method of section 2. The proof uses Lemmas 3.1–3.3 to show that the method is well defined and uses the observation (13) to show that (29) holds (unless $\nabla f(x_1^s) = 0$ for some $s \in T$ or $\liminf_{t\to\infty} f(x_1^t) = -\infty$). The latter in turn is used to show that $\{g^t\}_{t=0,1,...} \to 0$.

PROPOSITION 3.4. *The sequences* $\{(x_1^t, ..., x_{m+1}^t)\}_{t=0,1,...}$ *and* $\{\alpha^t\}_{t=0,1,...}$ *generated by the incremental gradient method (see* (2)–(3) *and* (4)–(11)*) are well defined.*

Moreover, either (i) $\nabla f(x_1^s) = 0$ *for some* $s \in T$ *or* (ii) $\liminf_{t\to\infty} f(x_1^t) = -\infty$ *or* (iii) $\{g^t\}_{t=0,1,\ldots} \to 0$ *and* $\{\nabla f(x_1^s)\}_{s\in T} \to 0$, *where* T *is the subsequence of* $\{1, 2, \ldots\}$ *specified in the method.*

Proof. We will show by induction on t that α^t is well defined for $t = 0, 1, \ldots$. Since $f(x_1^0) < \eta$, for $\alpha^0 \le \rho(1-\zeta)/\beta$, Lemma 3.2 shows that $x_1^0, \ldots, x_m^0$ are in $\mathcal{R}_\rho^\eta$ as is the line segment joining x_1^0 with x_1^1. Thus (5) holds and, by Lemma 3.3, (18) holds. Since $\epsilon_2 \frac{\zeta^m}{1-\zeta^m} < \eta - f(x_1^0)$, the latter implies that (6) holds for all α^0 sufficiently small, so α^0, being the largest element of $\{\epsilon_0, \omega\epsilon_0, \omega^2\epsilon_0, \ldots\}$ for which (6) holds, is well defined. Now assume that, for some $s \in T$, α^t is well defined for $t = 0, 1, \ldots, s-1$. If $\nabla f(x_1^s) = 0$, then we stop with case (i). Suppose instead that $\nabla f(x_1^s) \neq 0$ and let $h = \min\{t \in T : t > s\}$. We argue in the next two paragraphs that α^t is well defined for $t = s, \ldots, h-1$.

Since $g^t \to \nabla f(x_1^s)$ and β^t is bounded as $\alpha^t \to 0$ for $t = s, \ldots, h-1$, we have

$$\begin{aligned} &\alpha^t \le \frac{\rho(1-\zeta)}{\beta}, \quad \alpha^t \le \frac{(\delta_1 - \epsilon_1)\|g^t\|^2}{(\lambda(1.5 + 2\zeta\zeta_0^\infty) + \delta_3)(\beta^t)^2}, \\ &\alpha^t \le \frac{\epsilon_2}{\|g^t\|\bar{\beta}}, \quad \alpha^s \le \frac{\epsilon_3\|g^s\|}{\lambda\beta^s}, \quad t = s, \ldots, h-1, \end{aligned} \tag{25}$$

for any $\alpha^s = \cdots = \alpha^{h-1}$ sufficiently small. We claim that, whenever (25) holds, then

$$f(x_1^{t+1}) \le \eta, \qquad (x_2^t, \ldots, x_m^t, x_1^{t+1}) \in (\mathcal{R}_\rho^\eta)^m, \tag{26}$$

for $t = s-1, \ldots, h-1$. Clearly (26) holds for $t = s-1$. (This follows from (5)–(6) when $s = 1$ and follows from (7) with $t = s-1$ and (13) with $h = s$, with the latter implied by (8) with $h = s$, when $s > 1$.) Suppose that, for some $k \in \{s, \ldots, h-1\}$, the relation (26) holds for $t = s-1, \ldots, k-1$. Then, $(x_1^t, \ldots, x_m^t) \in (\mathcal{R}_\rho^\eta)^m$ for $t = 0, 1, \ldots, k-1$ (since (26) holds for $t = s-1, \ldots, k-1$ and (7) holds for $t = 0, 1, \ldots, s-1$ and $f(x_1^0) < \eta$) and, for each $t \in \{s, \ldots, k\}$, $f(x_1^t) \le \eta$. Since we also have $\alpha^t \le \rho(1-\zeta)/\beta$ (by (25)) for each $t \in \{s, \ldots, k\}$, Lemma 3.2 then yields that $x_1^t, \ldots, x_m^t$ and the line segment joining x_1^t with x_1^{t+1} all lie in $\mathcal{R}_\rho^\eta$, and, by Lemma 3.3, (17) holds. Using the second and third inequalities in (25) together with the fact $\alpha^t \le \alpha^\tau$, for $\tau = 0, 1, \ldots, t-1$, to bound the right-hand side of (17) yields

$$\begin{aligned} f(x_1^{t+1}) \le f(x_1^t) &- (\delta_2 + \epsilon_1)\,\alpha^t\|g^t\|^2 - \delta_3(\alpha^t\beta^t)^2 + \delta_2(1-\zeta)\sum_{\tau=1}^{t-1}\zeta_{t-\tau-1}^{t-\tau}\alpha^\tau\|g^\tau\|^2 \\ &+ \delta_3(1-\zeta)\sum_{\tau=0}^{t-1}\zeta_{t-\tau-1}^{t-\tau}(\alpha^\tau\beta^\tau)^2 + \epsilon_2\zeta^{tm} \end{aligned} \tag{27}$$

for $t = s, \ldots, k$. Also, we have that the inequality (8) holds for $h = s$. (For $s = 1$, this follows from (6); for $s > 1$, this follows from α^{s-1} being chosen in Step 1 such that (8) holds with $h = s$.) By using (12), we can rewrite this inequality equivalently as

$$\begin{aligned} f(x_1^s) \le \eta &- (\delta_2 + \epsilon_1)\sum_{\tau=1}^{s-1}\alpha^\tau\|g^\tau\|^2 - \delta_3\sum_{\tau=0}^{s-1}(\alpha^\tau\beta^\tau)^2 + \delta_2(1-\zeta)\sum_{\tau=1}^{s-2}\zeta_0^{s-\tau-1}\alpha^\tau\|g^\tau\|^2 \\ &+ \delta_3(1-\zeta)\sum_{\tau=0}^{s-2}\zeta_0^{s-\tau-1}(\alpha^\tau\beta^\tau)^2 - \frac{\epsilon_2\zeta^{sm}}{1-\zeta^m}. \end{aligned}$$

Summing (27) over all $t \in \{s, ..., k\}$ and then adding to it the above inequality, we obtain

$$\begin{aligned}
f(x_1^{k+1}) &\le \eta - (\delta_2+\epsilon_1)\sum_{\tau=1}^{k}\alpha^\tau\|g^\tau\|^2 - \delta_3\sum_{\tau=0}^{k}(\alpha^\tau\beta^\tau)^2 - \epsilon_2\left(\frac{\zeta^{sm}}{1-\zeta^m} - \sum_{t=s}^{k}\zeta^{tm}\right)\\
&\quad + \delta_2(1-\zeta)\left(\sum_{\tau=1}^{s-2}\zeta_0^{s-\tau-1}\alpha^\tau\|g^\tau\|^2 + \sum_{t=s}^{k}\sum_{\tau=1}^{t-1}\zeta_{t-\tau-1}^{t-\tau}\alpha^\tau\|g^\tau\|^2\right)\\
&\quad + \delta_3(1-\zeta)\left(\sum_{\tau=0}^{s-2}\zeta_0^{s-\tau-1}(\alpha^\tau\beta^\tau)^2 + \sum_{t=s}^{k}\sum_{\tau=0}^{t-1}\zeta_{t-\tau-1}^{t-\tau}(\alpha^\tau\beta^\tau)^2\right)\\
&= \eta - (\delta_2+\epsilon_1)\sum_{\tau=1}^{k}\alpha^\tau\|g^\tau\|^2 - \delta_3\sum_{\tau=0}^{k}(\alpha^\tau\beta^\tau)^2 - \epsilon_2\frac{\zeta^{(k+1)m}}{1-\zeta^m}\\
&\quad + \delta_2(1-\zeta)\sum_{\tau=1}^{k-1}\zeta_0^{k-\tau}\alpha^\tau\|g^\tau\|^2 + \delta_3(1-\zeta)\sum_{\tau=0}^{k-1}\zeta_0^{k-\tau}(\alpha^\tau\beta^\tau)^2 \qquad (28)\\
&\le \eta - (\delta_2+\epsilon_1)\sum_{\tau=1}^{k}\alpha^\tau\|g^\tau\|^2 - \delta_3\sum_{\tau=0}^{k}(\alpha^\tau\beta^\tau)^2 + \delta_2\sum_{\tau=1}^{k-1}\alpha^\tau\|g^\tau\|^2 + \delta_3\sum_{\tau=0}^{k-1}(\alpha^\tau\beta^\tau)^2,
\end{aligned}$$

where the equality follows from exchanging the order of summation and using the definition of ζ_i^j; the last inequality follows from the facts $\zeta_0^j \le \zeta_0^\infty = 1/(1-\zeta)$ for all $j > 0$. The right-hand side of the above inequality is less than η, so the claim (26) holds for $t = k$. By induction on k, it follows that (26) holds for $t = s-1, ..., h-1$.

Thus, if (25) holds, then (26) holds for $t = s-1, ..., h-1$, so that (7) holds. In addition, our argument showed that (28) holds for $k = s, ..., h-1$, so that, upon letting $k = h-1$ in (28) and using (12), we obtain (8). Also, we have from (24) with $t = s$ and the last inequality in (25) that

$$\|\nabla f(x_1^s) - g^s\| \le \lambda\alpha^s\beta^s \le \epsilon_3\|g^s\|,$$

so (9) holds. Thus (7)–(9) hold whenever $\alpha^s = ... = \alpha^{h-1}$ are sufficiently small. Then, $\alpha^s, ..., \alpha^{h-1}$, being the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, ..., \}$ such that (7)–(9) hold, are well defined. This completes the induction step and shows that α^t is well defined for all $t = 0, 1, ...$

There are three cases: either (i) $\nabla f(x_1^s) = 0$ for some $s \in T$ or else, since (8) holds and hence (13) holds for all $h \in T$, (ii) $\liminf_{t\to\infty} f(x_1^t) = -\infty$, or (iii)

$$\sum_{t=1}^{\infty}\alpha^t\|g^t\|^2 < \infty. \qquad (29)$$

Also, we have from $f(x_1^0) < \eta$ and (5) and (7) (with $h = \min\{t \in T : t > s\}$) for all $s \in T$ that

$$(x_1^t, ..., x_m^t) \in (\mathcal{R}_\rho^\eta)^m, \quad t = 0, 1, ... \qquad (30)$$

We claim that in case (iii), $\{g^t\} \to 0$. Since $\{\alpha^t\}$ is monotonically decreasing, either $\liminf_{t\to\infty}\alpha^t > 0$ or $\{\alpha^t\} \downarrow 0$. In the first case, (29) yields $\{g^t\} \to 0$. Consider now the second case. We showed earlier that, for each $s \in T$, (7)–(9) hold whenever

(25) holds, where $h = \min\{t \in T : t > s\}$. Then the choice of $\alpha^s = \cdots = \alpha^{h-1}$ being the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, ...,\}$ for which (7)–(9) hold, implies either $\alpha^s = \alpha^{s-1}$ or

$$\frac{\alpha^s}{\omega} > \min_{t=s,...,h-1} \left\{ \frac{\rho(1-\zeta)}{\beta}, \frac{(\delta_1 - \epsilon_1)\|g^t\|^2}{(\lambda(1.5 + 2\zeta\zeta_0^\infty) + \delta_3)(\beta^t)^2}, \frac{\epsilon_2}{\|g^t\|\bar{\beta}}, \frac{\epsilon_3\|g^s\|}{\lambda\beta^s} \right\}. \tag{31}$$

Since (30) holds, then for $t = 0, 1, ...$ we have from (10) and (14)–(15) that

$$\|g^t\| = \left\| \sum_{i=1}^m \nabla f_i(x_i^t) \right\| \le \sum_{i=1}^m \beta_i = \beta$$

and from (3) and (14)–(15) that

$$\sum_{i=1}^m \|d_i^t\| = \sum_{i=1}^m \left\| \sum_{k=0}^{tm+i-1} \zeta^k \nabla f_{i\ominus k} \left(x_{i\ominus k}^{\lfloor \frac{tm+i-1-k}{m} \rfloor} \right) \right\| \le \sum_{i=1}^m \sum_{k=0}^{tm+i-1} \zeta^k \beta_{i\ominus k} \le \frac{\beta}{1-\zeta},$$

so (10) yields $\beta^t \le \beta/(1-\zeta)$. Since $\{\alpha^s\} \downarrow 0$ so that (31) holds for all s in some subsequence of T, it follows that $g^t \to 0$ for t along some subsequence of $\{0, 1, ...\}$. We now argue that $g^t \to 0$ for t along the entire sequence $\{0, 1, ...\}$. Suppose this is not the case so there exists $\epsilon > 0$ such that $\|g^t\| > \epsilon$ for all t along some subsequence of $\{0, 1, ...\}$. The following argument is a modification of the proof of [13, Theorem 2.1]. Consider any t such that $\|g^t\| \ge \epsilon$. Since $\{\|g^t\|\}_{t=0,1,...}$ contains a subsequence that tends to zero, there exists a smallest integer $t' > t$ such that $\|g^{t'}\| < \epsilon/2$. Then,

$$\begin{aligned}
\frac{\epsilon}{2} &\le \|g^t\| - \|g^{t'}\| \\
&\le \|g^t - g^{t'}\| \\
&= \left\| \sum_{\tau=t}^{t'-1} \sum_{i=1}^m \left(\nabla f_i(x_i^{\tau+1}) - \nabla f_i(x_1^{\tau+1}) + \nabla f_i(x_{m+1}^\tau) - \nabla f_i(x_i^\tau) \right) \right\| \\
&\le \sum_{\tau=t}^{t'-1} \sum_{i=1}^m \lambda_i (\alpha^{\tau+1}\beta^{\tau+1} + \alpha^\tau \beta^\tau) \\
&\le \frac{2\lambda\beta}{1-\zeta} \sum_{\tau=t}^{t'-1} \alpha^\tau,
\end{aligned}$$

where the equality uses (2) and (10); the fourth inequality follows from (30) and Lemma 3.1; the last inequality follows from (15), $\beta^\tau \le \beta/(1-\zeta)$ for all τ and the monotone decreasing property of $\{\alpha^\tau\}_{\tau=0,1,...}$. We also have that $\|g^\tau\| \ge \epsilon/2$ for $\tau = t, ..., t' - 1$, which together with the above relation yield

$$\sum_{\tau=t}^{t'-1} \alpha^\tau \|g^\tau\|^2 \ge \frac{\epsilon^2}{4} \sum_{\tau=t}^{t'-1} \alpha^\tau \ge \frac{\epsilon^2}{4} \frac{\epsilon(1-\zeta)}{4\lambda\beta}.$$

Since the number of such t is infinite, it follows that $\sum_{\tau=1}^\infty \alpha^\tau \|g^\tau\|^2 = \infty$, a contradiction of (29).

Since $\{g^t\} \to 0$ and (9) holds for all $s \in T$, it follows that $\{\nabla f(x_1^s)\}_{s\in T} \to 0$. □

4. Convergence rate and stepsize analysis. In this section, we show that under a growth assumption on $\nabla f_1, ..., \nabla f_m$ (see Assumption B below), the incremental gradient method either is linearly convergent in some sense or has its stepsize bounded away from zero (see Proposition 4.2). This result gives an explanation of the observed behavior that on some problems, the stepsize remains bounded away from zero (see the numerical experience reported in section 6).

To establish our result, we first need the following technical lemma.

LEMMA 4.1. *For any $\alpha^t \in (0, \alpha^0]$ and any $x_1^t, ..., x_{m+1}^t$ in $\mathcal{R}_\rho^\eta$ satisfying (2), with $d_1^t, ..., d_m^t$ given by (3), for $t = 0, 1, ...$, we have*

$$\|d_i^t\| \le \sum_{\tau=0}^{t-1} \mu_{t-\tau}^{t-\tau+1} h^\tau + (1 + \cdots + \mu^{i-1}) h^t, \tag{32}$$

for $i = 1, ..., m$ and $t = 0, 1, ...$, where we let

$$h^t := \max_{i=1,...,m} \|\nabla f_i(x_1^t)\|, \quad t = 0, 1, ..., \quad \mu := \left((m-1)\alpha^0 \max_{i=1,...,m} \lambda_i + \zeta \right)^{1/m}. \tag{33}$$

Proof. Clearly, (32) holds for $t = 0$ and $i = 1$. Suppose that, for some $s \ge 0$ and some $1 \le j \le m$, (32) holds for $i = 1, ..., m$ if $t < s$ and for $i = 1, ..., j$ if $t = s$. First, consider the case $j = m$. Then, by (3) with $t = s + 1$ and (32) with $t = s$ and $i = m$,

$$\|d_1^{s+1}\| = \|\nabla f_1(x_1^{s+1}) + \zeta d_m^s\| \le \|\nabla f_1(x_1^{s+1})\| + \zeta \|d_m^s\| \le h^{s+1} + \zeta \left(\sum_{\tau=0}^{s} \mu_{s-\tau}^{s-\tau+1} h^\tau \right).$$

Since $\zeta \le \mu^m$, this implies (32) holds for $t = s + 1$ and $i = 1$. Second, consider the case $j < m$. Then, by (3) for $t = s$ and $i = j + 1$,

$$\begin{aligned}
\|d_{j+1}^s\| &= \|\nabla f_{j+1}(x_{j+1}^t) + \zeta d_j^t\| \\
&\le \|\nabla f_{j+1}(x_{j+1}^t) - \nabla f_{j+1}(x_1^t)\| + \|\nabla f_{j+1}(x_1^t)\| + \zeta \|d_j^t\| \\
&\le \alpha^t \lambda_{j+1} (\|d_j^t\| + \cdots + \|d_1^t\|) + \|\nabla f_{j+1}(x_1^t)\| + \zeta \|d_j^t\| \\
&\le j\alpha^t \lambda_{j+1} \left(\sum_{\tau=0}^{t-1} \mu_{t-\tau-1}^{t-\tau} h^\tau + (1 + \cdots + \mu^{j-1}) h^t \right) \\
&\quad + h^t + \zeta \left(\sum_{\tau=0}^{t-1} \mu_{t-\tau-1}^{t-\tau} h^\tau + (1 + \cdots + \mu^{j-1}) h^t \right) \\
&= (j\alpha^t \lambda_{j+1} + \zeta) \sum_{\tau=0}^{t-1} \mu_{t-\tau-1}^{t-\tau} h^\tau + (j\alpha^t \lambda_{j+1} + \zeta)(1 + \cdots + \mu^{j-1}) h^t + h^t \\
&\le \mu^m \sum_{\tau=0}^{t-1} \mu_{t-\tau-1}^{t-\tau} h^\tau + (1 + \cdots + \mu^j) h^t,
\end{aligned}$$

where the second inequality follows from $x_{j+1}^t, x_1^t \in \mathcal{R}_\rho^\eta$ and the Lipschitz continuity of ∇f_{j+1} on $\mathcal{R}_\rho^\eta$ (with constant λ_{j+1}) and (2), the third inequality follows from (32) with $i = 1, ..., j$ and (33). This implies (32) holds for $t = s$ and $i = j + 1$. Thus, by induction on t and i, (32) holds for all $i = 1, ..., m$ and all $t = 0, 1, ...$. ∎

Consider the following growth assumption on $f_1, ..., f_m$.

Assumption B. There exists $c_1 > 0$ such that

$$\max_{i=1,...,m} \|\nabla f_i(x)\| \leq c_1 \|\nabla f(x)\| \quad \forall x \in \mathcal{R}_\rho^\eta.$$

Assumption B roughly says that the size of $\nabla f_1, ..., \nabla f_m$ should grow no faster than linearly with the size of ∇f. In particular, this assumption requires that $\nabla f_1(x) = \cdots = \nabla f_m(x) = 0$ at any stationary point x of f. This requirement, though restrictive, is not entirely unrealistic for certain applications. For example, for the application of neural network training, this requirement amounts to being able to train the neural network to achieve zero output error on the learning examples. In fact, it is possible for this requirement to fail to hold and still have the stepsize bounded away from zero. Consider the example of $n = 1, m = 2$ and $f_1(x) = x$, $f_2(x) = -x$. Then, $f \equiv 0$ and Assumption A holds with $\eta = \rho = \infty$ and $\lambda_1 = \lambda_2 = 0$. Upon applying the incremental gradient method with, say, $\zeta = 0$ and any choice of x_1^0, ω, T, and $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3$ satisfying $\epsilon_1 < 1$, we find that $\delta_2 = \delta_3 = g^t = p^t = 0$ and, in particular, $\alpha^t = \epsilon_0$ for all $t = 0, 1, ...$ In contrast, the other stepsize rules mentioned in section 1 would require the stepsize to tend to zero on this example. (This example is degenerate in the sense that every $x \in \Re$ is a stationary point of f. However, it can be easily modified so that x_1^0 is not a stationary point, etc.) In general, if the iterates are in a region where $f_1, ..., f_m$ are nearly linear, then the stepsize will tend not to decrease.

By using Lemma 4.1 and the fact (see the proof of Proposition 3.4) that (31) holds whenever $\alpha^s \neq \alpha^{s-1}$, we have the following convergence rate and stepsize result for the incremental gradient method. The result roughly says that, under Assumption B, either h^t given by (33) tends to zero linearly in some sense or α^t is bounded away from zero. The proof of this uses the idea that if h^t does not tend to zero linearly, then neither does $\|\nabla f(x_1^t)\|$ (by Assumption B) from which it can be shown that $\|d_i^t\| = O(\|\nabla f(x_1^t)\|)$ (see (35)). This in turn can be used to show that the right-hand side of (31) is bounded away from zero.

PROPOSITION 4.2. *Assume Assumption B (in addition to Assumption A) is satisfied and let* $\{(x_1^t, ..., x_{m+1}^t)\}_{t=0,1,...}$ *and* $\{\alpha^t\}_{t=0,1,...}$ *be generated by the incremental gradient method* (2)–(3) *and* (4)–(13) *(which, by Prop.* 3.4, *are well defined) with* α^0 *chosen sufficiently small so that* μ *given by* (33) *is less than* 1. *If there exists a* $c_2 \geq 1$ *and* $\sigma \in (\mu, 1)$ *such that*

$$\sigma_{t-\tau}^{t-\tau+1} h^\tau \leq c_2 h^t, \qquad \tau = 0, ..., t-1, \tag{34}$$

for $t = 1, 2, ...$, *where* h^t *is given by* (33), *then* $\liminf_{t\to\infty} \alpha^t > 0$.

Proof. Fix any $t \in \{1, 2, ..\}$. Since (30) holds, then Lemma 4.1 and $c_2 \geq 1$ and $\sigma < 1$ show that, for each $i = 1, ..., m$, d_i^t given by (3) satisfies

$$\begin{aligned}
\|d_i^t\| &\leq \sum_{\tau=0}^{t-1} \mu_{t-\tau}^{t-\tau+1} h^\tau + \left(1 + \cdots + \left(\frac{\mu}{\sigma}\right)^{m-1}\right) c_2 h^t \\
&\leq \sum_{\tau=0}^{t-1} \mu_{t-\tau}^{t-\tau+1} \left(\frac{c_2 h^t}{\sigma_{t-\tau}^{t-\tau+1}}\right) + \left(1 + \cdots + \left(\frac{\mu}{\sigma}\right)^{m-1}\right) c_2 h^t \\
&\leq \sum_{\tau=0}^{t-1} \left(\frac{\mu}{\sigma}\right)_{t-\tau}^{t-\tau+1} c_2 h^t + \left(1 + \cdots + \left(\frac{\mu}{\sigma}\right)^{m-1}\right) c_2 h^t
\end{aligned}$$

$$= \sum_{\tau=0}^{t} \Big(\frac{\mu}{\sigma}\Big)_{t-\tau}^{t-\tau+1} c_2 h^t$$

$$\leq \frac{c_2 h^t}{1-\mu/\sigma}$$

$$\leq c_3 \|\nabla f(x_1^t)\|, \tag{35}$$

where the second inequality follows from (34); the third inequality follows from the fact $(a_1+\cdots+a_m)/(b_1+\cdots+b_m) \leq \max_{i=1,\ldots,m} a_i/b_i \leq (a_1/b_1+\cdots+a_m/b_m)$; the last inequality follows by using Assumption B and letting $c_3 := \frac{c_2}{1-\mu/\sigma} c_1$. Then (10) and (15) yield

$$\|g^t - \nabla f(x_1^t)\| \leq \sum_{i=1}^{m} \|\nabla f_i(x_i^t) - \nabla f_i(x_1^t)\|$$

$$\leq \sum_{i=1}^{m} \lambda_i \|x_1^t - x_i^t\|$$

$$= \sum_{i=1}^{m} \lambda_i \alpha^t \|d_1^t + \cdots + d_{i-1}^t\|$$

$$\leq \lambda \alpha^t m c_3 \|\nabla f(x_1^t)\|, \tag{36}$$

where d_i^t is given by (3), the second inequality follows from ∇f_i being Lipschitz continuous on $\mathcal{R}_\rho^\eta$ (with constant λ_i) and the equality follows from (2). Thus,

$$\|g^t\| \geq \|\nabla f(x_1^t)\| - \|g^t - \nabla f(x_1^t)\| \geq (1 - \lambda \alpha^t m c_3) \|\nabla f(x_1^t)\|$$

and, by (10) and (35)–(36),

$$\beta^t = \max\left\{ \|g^t\|, \sum_{i=1}^{m} \|d_i^t\| \right\}$$

$$\leq \max\left\{ \|g^t - \nabla f(x_1^t)\| + \|\nabla f(x_1^t)\|, m c_3 \|\nabla f(x_1^t)\| \right\}$$

$$\leq \max\left\{ \lambda \alpha^t m c_3 + 1, m c_3 \right\} \|\nabla f(x_1^t)\|.$$

Thus, if $\{\alpha^s\} \downarrow 0$, then (31) must hold for all s along some subsequence of T. On the other hand, the above two inequalities and $\|g^t\| \leq \beta$ (which hold for all t) show that the right-hand side of (31) is bounded away from zero, which implies α^s is bounded away from zero for all s along this subsequence, a contradiction to $\{\alpha^s\} \downarrow 0$. □

Note that, for all t, $h^t \geq \frac{1}{n} \sum_{i=1}^{m} \|\nabla f_i(x_1)\| \geq \frac{1}{n} \|\nabla f(x_1^t)\|$ while, by $x_1^t \in \mathcal{R}_\rho^\eta$ and Assumption B, $h^t \leq c_1 \|\nabla f(x_1^t)\|$. Thus, under Assumption B, (34) is equivalent to

$$\sigma_{t-\tau}^{t-\tau+1} \|\nabla f(x_1^\tau)\| \leq c_2' \|\nabla f(x_1^t)\|, \quad \tau = 0, 1, ..., t-1,$$

for some constant c_2'. In the very special case where $f_1 = \cdots = f_m$ (so Assumption B holds trivially), the above proof can be modified to show that α^t generated by the incremental gradient method with $\zeta = 0$ (i.e., steepest descent) is always bounded away from zero.

5. Extension to constrained problems. In this section, we consider an extension of the incremental gradient method of section 2 to the problem of minimizing, over a nonempty closed convex set $\mathcal{X}$ of $\Re^n$, the function f given by (1). Such constrained problems arise in neural network training where bounds are placed on the weights of the neural network and which corresponds to $\mathcal{X}$ being a box. Due to the presence of the constraint set $\mathcal{X}$, the formula for updating x_i^t and d_i^t need to be modified much as in [12]. We show that, analogous to Proposition 3.4, this extended method has desirable global convergence properties (see Proposition 5.4).

Consider the following iterative method for solving the preceding problem whereby, for a given $x_1^0 \in \mathcal{X}$, we generate a sequence $(x_1^t, ..., x_{m+1}^t)$, $t = 0, 1, ...$, according to

$$\begin{aligned} x_{i+1}^t &:= x_i^t - \alpha^t d_i^t, \quad i = 1, ..., m, \\ x_1^{t+1} &:= [x_{m+1}^t]^+, \end{aligned} \tag{37}$$

where α^t is a positive scalar and

$$d_i^t := \begin{cases} \nabla f_i(x_i^t) + \zeta \nabla f_m(x_m^{t-1}) & \text{if } i = 1, \\ \nabla f_i(x_i^t) + \zeta \nabla f_{i-1}(x_{i-1}^t) & \text{if } i > 1 \end{cases} \tag{38}$$

with $x_m^{-1} = x_1^0$ given. (Here, $[\cdot]^+$ denotes the orthogonal projection operator onto $\mathcal{X}$. This operator can be evaluated fairly easily if $\mathcal{X}$ is either a box or a Euclidean sphere or simplex.) In the context of neural network training, the projection in (37) corresponds to the oft-used practice of truncating the weights of the neural network at their respective bounds. Note that (38) sets d_i^t to be a weighted sum of the two most recently computed gradients. This contrasts with (3), which sets d_i^t to be a weighted sum of *all* previously computed gradients. Thus, while (37) reduces to (2) when $\mathcal{X} = \Re^n$, the formula (38) does not reduce to (3) when $\mathcal{X} = \Re^n$, so (37)–(38) is a different method from (2)–(3). Our results can be extended to the case where d_i^t is a weighted sum of the K ($K \geq 1$ and constant) most recently computed gradients though, for simplicity, we will not consider this more general case here. It is an open question whether our results can be extended to the case where (38) is replaced by (3). Also, we note that in the case where $\zeta = 0$, the method (37)–(38) reduces to the approximate gradient-projection method studied in [12]. Our preliminary numerical experience suggests that $\zeta > 0$ (e.g., $\zeta = .1$) is typically preferable.

As with the incremental gradient method of section 2, we propose a new rule for choosing the stepsize α^t adaptively. To describe this new stepsize rule, we make the following assumption, analogous to Assumption A, about $f, f_1, ..., f_m$, $\mathcal{X}$ and the initial iterate x_1^0.

Assumption C. There exist scalars $\eta > f(x_1^0)$ and $\rho > 0$ such that, for $i = 1, ..., m$, ∇f_i is bounded and Lipschitz continuous (with some constant λ_i) on the set

$$\mathcal{X}_\rho^\eta := \{ \, x \in \mathcal{X} \; : \; f(x) \leq \eta \, \} + \rho \mathcal{B},$$

where $\mathcal{B} = \{ \, x \in \Re^n \; : \; \|x\| \leq 1 \, \}$.

Assumption C is quite mild and, in particular, is satisfied when $f_1, ..., f_m$ are twice differentiable and the level set $\{ \, x \in \mathcal{X} \; : \; f(x) \leq \eta \, \}$ is bounded for some $\eta > f(x_1^0)$.

The new stepsize rule for choosing α^t, analogous to the one of section 2, depends on η, ρ and $\lambda_1, ..., \lambda_m$ and periodically checks if a certain descent condition is satisfied since the previous check was made and, if not, decreases the stepsize and restarts

the method from when the previous check was made. Below, we formally state the method (37)–(38) using this stepsize rule. We will call this method the *incremental gradient-projection* method.

Incremental gradient-projection method (with 1-memory momentum term). Choose any $x_1^0 \in \mathcal{X}$ such that Assumption C holds for some η, ρ and $\lambda_1, ..., \lambda_m$. Choose any $\zeta \in (0, \infty)$. Choose any $\omega \in (0,1)$ and any subsequence T of $\{1, 2, ...\}$ containing 1. Choose any positive scalars $\epsilon_0, \epsilon_1, \epsilon_2$ satisfying $\epsilon_1 < 1 + \zeta$.

Step 0. Let α^0 be the largest element of $\{\epsilon_0, \omega\epsilon_0, \omega^2\epsilon_0, ...\}$ for which $x_2^0, ..., x_m^0, x_1^1$ given by (37)–(38) with $t = 0$ satisfy the following two conditions:

$$(x_2^0, ..., x_m^0, x_1^1) \in (\mathcal{X}_\rho^\eta)^m, \tag{39}$$

$$f(x_1^1) \le \eta - \lambda_m \zeta (\alpha^0 \beta^0)^2, \tag{40}$$

where β^0 is given by (10).

Step 1. For each $s \in T$, let $h = \min\{t \in T : t > s\}$ and we generate $(x_1^h, d_m^{h-1}, \alpha^{h-1})$ from $(x_1^s, d_m^{s-1}, \alpha^{s-1})$ as follows: If $r^s = 0$, we stop. Else, we let α^{h-1} be the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, \omega^2\alpha^{s-1}, ...\}$ for which $x_2^t, ..., x_m^t, x_1^{t+1}$ given by (37)–(38) with $\alpha^t = \alpha^{h-1}$, $t = s, ..., h-1$, satisfy the following three conditions:

$$(x_2^t, ..., x_m^t, x_1^{t+1}) \in (\mathcal{X}_\rho^\eta)^m, \quad t = s, ..., h-1, \tag{41}$$

$$f(x_1^h) \le \eta - \epsilon_1 \sum_{\tau=1}^{h-1} \alpha^\tau \|\hat{r}^\tau\|^2 - \lambda_m \zeta (\alpha^{h-1}\beta^{h-1})^2, \tag{42}$$

$$\|r^s - \hat{r}^s\| \le \epsilon_2 \|\hat{r}^s\|, \tag{43}$$

where, for $t = s, ..., h-1$, we define

$$r^t := [x_1^t - \nabla f(x_1^t)]^+ - x_1^t, \qquad \hat{r}^t := [x_1^t - g^t]^+ - x_1^t, \tag{44}$$

with g^t and β^t given by (10).

Like the stepsize rule of section 2, the above stepsize rule checks at each iteration $s \in T$ whether the current stepsize is acceptable (i.e., satisfies (41)–(43)) for all iterations between s and the next element h of T and, if not, it decreases the stepsize by ω and restarts the method from iteration s. Note that the conditions (40) and (42) are much simpler than their counterpart (6) and (8) of section 2. This is because here d_i^t depends only on the two most recently computed gradients. The quantity $\hat{r}^t$ may be viewed as an approximation to r^t, the "natural residual" at x_1^t.

To establish the global convergence of the incremental gradient-projection method, we first need the following three technical lemmas analogous to Lemmas 3.1–3.3. We assume throughout that Assumption C holds.

LEMMA 5.1. *For any $t \in \{0, 1, ...\}$, any $\alpha^t > 0$, and any $x_1^t, ..., x_m^t, x_1^{t+1}$ satisfying* (37)*, we have, for $i = 1, ..., m$, that*

$$\|\nabla f_i(x_1^t) - \nabla f_i(x_i^t)\| \le \lambda_i \alpha^t \beta^t$$

whenever $x_1^t, x_i^t \in \mathcal{X}_\rho^\eta$ and that

$$\|\nabla f_i(x_1^{t+1}) - \nabla f_i(x_i^t)\| \le 2\lambda_i \alpha^t \beta^t$$

whenever $x_i^t, x_1^{t+1} \in \mathcal{X}_\rho^\eta$, where β^t is given by (10).

Proof. We have from (37) and (10) that for any $i \in \{1, ..., m+1\}$,

$$\|x_1^t - x_i^t\| = \alpha^t \left\| \sum_{l=1}^{i-1} d_l^t \right\| \leq \alpha^t \beta^t.$$

By a similar reasoning as above, we have $\|x_{m+1}^t - x_i^t\| \leq \alpha^t \beta^t$ which, together with the above inequality with $i = m+1$, yields

$$\begin{aligned} \|x_1^{t+1} - x_i^t\| &\leq \|x_1^{t+1} - x_{m+1}^t\| + \|x_{m+1}^t - x_i^t\| \\ &\leq \|x_1^t - x_{m+1}^t\| + \|x_{m+1}^t - x_i^t\| \\ &\leq \alpha^t \beta^t + \alpha^t \beta^t, \end{aligned}$$

where the second inequality follows from the observation that x_1^{t+1} is the point in $\mathcal{X}$ nearest in Euclidean distance to x_{m+1}^t and $x_1^t \in \mathcal{X}$. The above inequalities, together with ∇f_i being Lipschitz continuous (with constant λ_i) on $\mathcal{X}_\rho^\eta$ for $i = 1, ..., m$, yield the desired results. □

Under Assumption C, there exist positive scalars $\beta_1, ..., \beta_m$ such that

$$\|\nabla f_i(x)\| \leq \beta_i \qquad \forall x \in \mathcal{X}_\rho^\eta, \qquad i = 1, ..., m. \tag{45}$$

Let β and λ be given by (15). Also, let

$$\delta_0 := \rho/(2\beta(1+\zeta)). \tag{46}$$

LEMMA 5.2. *For any $t \in \{0, 1, ...\}$, any $\alpha^t \in (0, \delta_0]$, and any $x_m^{t-1}, x_1^t, ..., x_m^t, x_1^{t+1}$ satisfying (37)–(38) and such that $x_m^{t-1} \in \mathcal{X}_\rho^\eta$, $x_1^t \in \mathcal{X}$ and $f(x_1^t) \leq \eta$, we have that $x_1^t, ..., x_m^t$ are in $\mathcal{X}_\rho^\eta$, as is the line segment joining x_1^t with x_1^{t+1}.*

Proof. We claim that

$$\|x_l^t - x_1^t\| \leq \frac{\rho}{2\beta(1+\zeta)} \sum_{k=1}^{l-1} (\beta_k + \zeta \beta_{k\ominus 1}) \tag{47}$$

for $l = 1, ..., m+1$. We prove this by induction on l. Clearly, (47) holds for $l = 1$. Suppose (47) holds for $l = 1, ..., i$ for some $i \in \{1, ..., m\}$ and we show below that it also holds for $l = i+1$. Since (47) holds for $l = 1, ..., i$ and the right-hand side of (47) is bounded above by $\rho/2$, we have (cf. $f(x_1^t) \leq \eta$ and $x_1^t \in \mathcal{X}$) $x_1^t, ..., x_i^t \in \mathcal{X}_\rho^\eta$, as well as $x_m^{t-1} \in \mathcal{X}_\rho^\eta$. Then, (37)–(38) and (45) and $\alpha^t \leq \delta_0$ yield

$$\begin{aligned} \|x_{i+1}^t - x_i^t\| = \alpha^t \left\| \nabla f_i(x_i^t) + \zeta \nabla f_{i\ominus 1} \left(x_{i\ominus 1}^{\lfloor \frac{tm+i-2}{m} \rfloor} \right) \right\| &\leq \alpha^t (\beta_i + \zeta \beta_{i\ominus 1}) \\ &\leq \frac{\rho}{2\beta(1+\zeta)} (\beta_i + \zeta \beta_{i\ominus 1}). \end{aligned}$$

Since (47) holds for $l = i$, this shows that (47) holds for $l = i+1$.

The claim (47) together with $f(x_1^t) \leq \eta$ and $x_1^t \in \mathcal{X}$ implies that $x_l^t \in \mathcal{X}_\rho^\eta$ for $l = 1, ..., m+1$. Also, since x_1^{t+1} is the point in $\mathcal{X}$ nearest in Euclidean distance to x_{m+1}^t (see (37)) and $x_1^t \in \mathcal{X}$, we have

$$\|x_{m+1}^t - x_1^{t+1}\| \leq \|x_{m+1}^t - x_1^t\| \leq \rho/2,$$

where the last inequality follows from (47) with $l = m + 1$ and then bounding the right-hand side of (47) from above by $\rho/2$. Thus,

$$\|x_1^t - x_1^{t+1}\| \le \|x_1^t - x_{m+1}^t\| + \|x_{m+1}^t - x_1^{t+1}\| \le \rho$$

and, hence, (cf. $f(x_1^t) \le \eta$ and $x_1^t \in \mathcal{X}$) the line segment joining x_1^t with x_1^{t+1} lies in $\mathcal{X}_\rho^\eta$. □

By using Lemma 5.1, we obtain the third lemma estimating the decrease in f-value per iteration of the incremental gradient-projection method.

LEMMA 5.3. *For any $t \in \{1, 2, ...\}$, any $\alpha^\tau \in (0, 1/(1+\zeta)]$, and any $x_m^{\tau-1}, x_1^\tau, ..., x_m^\tau$, $x_1^{\tau+1}$ satisfying (37)–(38) for $\tau = t-1, t$, and such that $x_1^t \in \mathcal{X}$ and both x_m^{t-1}, x_m^t and the line segment joining x_1^t with x_1^{t+1} lie in $\mathcal{X}_\rho^\eta$, we have*

$$f(x_1^{t+1}) \le f(x_1^t) - (1+\zeta)\alpha^t\|\hat{r}^t\|^2 + (1.5\lambda + 2\lambda_m\zeta)(\alpha^t\beta^t)^2 + \lambda_m\zeta(\alpha^{t-1}\beta^{t-1})^2, \tag{48}$$

where β^t is given by (10) and $\hat{r}^t$ is given by (44). Similarly, for any $\alpha^0 \in (0, 1/(1+\zeta)]$ and any $x_m^{-1}, x_1^0, ..., x_m^0, x_1^1$ satisfying (37)–(38) (with $t = 0$) and such that $x_m^{-1} = x_1^0 \in \mathcal{X}$ and both x_m^0 and the line segment joining x_1^0 with x_1^1 lie in $\mathcal{X}_\rho^\eta$, we have

$$f(x_1^1) \le f(x_1^0) - (1+\zeta)\alpha^0\|\hat{r}^0\|^2 + (1.5\lambda + \lambda_m\zeta)(\alpha^0\beta^0)^2. \tag{49}$$

Proof. Fix any $t \in \{1, 2, ...\}$. Since the line segment joining x_1^t with x_1^{t+1} lies in $\mathcal{X}_\rho^\eta$ and, by $\nabla f = \nabla f_1 + \cdots + \nabla f_m$ (cf. (1)) and Assumption C, ∇f is Lipschitz continuous (with constant λ given by (15)) on $\mathcal{X}_\rho^\eta$, we obtain from the intermediate value theorem (see [4, p. 639]) that

$$f(x_1^{t+1}) \;\le\; f(x_1^t) + \langle \nabla f(x_1^t), x_1^{t+1} - x_1^t\rangle + ..5\lambda\|x_1^{t+1} - x_1^t\|^2. \tag{50}$$

Using (37) and $x_1^t \in \mathcal{X}$, we bound the rightmost term in (50) as follows:

$$\|x_1^{t+1} - x_1^t\| = \left\| \left[x_1^t - \alpha^t \sum_{i=1}^m d_i^t \right]^+ - [x_1^t]^+ \right\| \le \alpha^t \left\| \sum_{i=1}^m d_i^t \right\| \le \alpha^t\beta^t, \tag{51}$$

where the first inequality follows from the nonexpansive property of the projection operator $[\cdot]^+$, and the last inequality follows from (10). The second term on the right-hand side of (50) can be bounded as follows:

$$\begin{aligned}
&\langle \nabla f(x_1^t), x_1^{t+1} - x_1^t\rangle \\
&= \langle \nabla f(x_1^t) - g^t, x_1^{t+1} - x_1^t\rangle + \langle g^t, [x_1^t - \alpha^t(1+\zeta)g^t]^+ - x_1^t\rangle \\
&\qquad\qquad + \langle g^t, x_1^{t+1} - [x_1^t - \alpha^t(1+\zeta)g^t]^+\rangle \\
&\le \|\nabla f(x_1^t) - g^t\|\|x_1^{t+1} - x_1^t\| - \frac{1}{\alpha^t(1+\zeta)}\|[x_1^t - \alpha^t(1+\zeta)g^t]^+ - x_1^t\|^2 \\
&+ \|g^t\|\|x_1^{t+1} - [x_1^t - \alpha^t(1+\zeta)g^t]^+\| \\
&\le \|\nabla f(x_1^t) - g^t\|\|x_1^{t+1} - x_1^t\| - \alpha^t(1+\zeta)\|\hat{r}^t\|^2 + \|g^t\|\|x_1^{t+1} - [x_1^t - \alpha^t(1+\zeta)g^t]^+\| \\
&\le \|\nabla f(x_1^t) - g^t\|\|x_1^{t+1} - x_1^t\| - \alpha^t(1+\zeta)\|\hat{r}^t\|^2 + \alpha^t\|g^t\| \left\| \sum_{i=1}^m d_i^t - (1+\zeta)g^t \right\|,
\end{aligned}$$

where the first inequality follows from the Cauchy–Schwarz inequality and the following well-known property of $[\cdot]^+$:

$$\langle x-y,[y]^+-x\rangle \le -\|[y]^+-x\|^2 \qquad \forall x\in\mathcal{X}\ \forall y\in\Re^n;$$

the second inequality follows from $\alpha^t(1+\zeta)\le 1$ and $\|[x-\gamma d]^+-x\|\ge\gamma\|[x-d]^+-x\|$ for all $\gamma\in[0,1]$ (see Lemma 1 in [7]) and (44); the last inequality follows from $x_1^{t+1}=[x_1^t-\alpha^t\sum_{i=1}^m d_i^t]^+$ (see (37)) and the nonexpansive property of $[\cdot]^+$. Also, we have from (10) and (38) that

$$\begin{aligned}\left\|\sum_{i=1}^m d_i^t-(1+\zeta)g^t\right\| &= \zeta\left\|\nabla f_m(x_m^{t-1})-\nabla f_m(x_m^t)\right\|\\ &\le \zeta\left(\left\|\nabla f_m(x_m^{t-1})-\nabla f_m(x_1^t)\right\|+\left\|\nabla f_m(x_1^t)-\nabla f_m(x_m^t)\right\|\right)\\ &\le \zeta\lambda_m(2\alpha^{t-1}\beta^{t-1}+\alpha^t\beta^t),\end{aligned}$$

where the last inequality follows from $x_m^{t-1},x_1^t,x_m^t\in\mathcal{X}_\rho^\eta$ and Lemma 5.1. By a similar argument, we have that (24) holds. Combining the above two inequalities with (24) and $\|g^t\|\le\beta^t$ (see (10)) yields

$$\begin{aligned}\langle\nabla f(x_1^t),x_1^{t+1}-x_1^t\rangle &\le \lambda(\alpha^t\beta^t)^2-\alpha^t(1+\zeta)\|\hat r^t\|^2+\alpha^t\beta^t\zeta\lambda_m(2\alpha^{t-1}\beta^{t-1}+\alpha^t\beta^t)\\ &\le \lambda(\alpha^t\beta^t)^2-\alpha^t(1+\zeta)\|\hat r^t\|^2+\zeta\lambda_m((\alpha^{t-1}\beta^{t-1})^2+2(\alpha^t\beta^t)^2),\end{aligned}$$

where the last inequality uses $ab\le .5(a^2+b^2)$. This together with (50)–(51) proves (48). The proof of (49) is very similar. □

Below we state and prove the global convergence result for the incremental gradient-projection method. The proof uses Lemmas 5.1–5.3.

PROPOSITION 5.4. *The sequences* $\{(x_1^t,...,x_{m+1}^t)\}_{t=0,1,...}$ *and* $\{\alpha^t\}_{t=0,1,...}$, *generated by the incremental gradient-projection method (see* (37)–(38) *and* (39)–(44)*) are well defined. Moreover, either* (i) $r^s=0$ *for some* $s\in T$ *or* (ii) $\liminf_{t\to\infty} f(x_1^t)=-\infty$ *or* (iii) $\{\hat r^t\}\to 0$ *and* $\{r^t\}_{t\in T}\to 0$, *where* T *is the subsequence of* $\{1,2,...\}$ *specified in the method.*

Proof. We show by induction on t that α^t is well defined for $t=0,1,....$ Since $x_m^{-1}=x_1^0\in\mathcal{X}$ and $f(x_1^0)<\eta$, for $\alpha^0\le\min\{\delta_0,1/(1+\zeta)\}$, Lemma 5.2 shows that $x_1^0,...,x_m^0$ are in $\mathcal{X}_\rho^\eta$ as is the line segment joining x_1^0 with x_1^1. Thus (39) holds and, by Lemma 5.3, (49) holds. Since $f(x_1^0)<\eta$, the latter implies that (40) holds for all α^0 sufficiently small, so α^0, being the largest element of $\{\epsilon_0,\omega\epsilon_0,\omega^2\epsilon_0,...\}$ for which (40) holds, is well defined. Now assume that, for some $s\in T$, α^t is well defined for $t=0,1,...,s-1$. If $\nabla f(x_1^s)=0$, then we stop with case (i). Suppose instead $\nabla f(x_1^s)\ne 0$ and let $h=\min\{t\in T: t>s\}$. We argue in the next two paragraphs that α^t is well defined for $t=s,...,h-1$.

Since $g^t\to\nabla f(x_1^s)$ and β^t is bounded as $\alpha^t\to 0$ for $t=s,...,h-1$, we have

(52)

$$\alpha^t\le\min\{\delta_0,1/(1+\zeta)\},\quad \alpha^t\le\frac{(1+\zeta-\epsilon_1)\|\hat r^t\|^2}{(1.5\lambda+3\lambda_m\zeta)(\beta^t)^2},\quad \alpha^s\le\frac{\epsilon_2\|\hat r^s\|}{\lambda\beta^s},\quad t=s,...,h-1,$$

for any $\alpha^s=\cdots=\alpha^{h-1}$ sufficiently small. We claim that, whenever (52) holds, then

$$f(x_1^{t+1})\le\eta,\qquad (x_2^t,...,x_m^t,x_1^{t+1})\in(\mathcal{X}_\rho^\eta)^m, \tag{53}$$

for $t = s-1, ..., h-1$. Clearly (53) holds for $t = s-1$. (This follows from (39)–(40) when $s = 1$ and follows from (41) with $t = s-1$ and (42) with $h = s$ when $s > 1$.) Suppose that, for some $k \in \{s, ..., h-1\}$, the relation (53) holds for $t = s-1, ..., k-1$. From (37), we also have $x_1^{t+1} \in \mathcal{X}$ for $t = s-1, ..., k-1$. Then, for each $t \in \{s, ..., k\}$, we have $f(x_1^t) \leq \eta$, $x_m^{t-1} \in \mathcal{X}_\rho^\eta$ and $x_1^t \in \mathcal{X}$, which together with Lemma 5.2 and $\alpha^t \leq \delta_0$ (see (52)) imply that $x_1^t, ..., x_m^t$ and the line segment joining x_1^t with x_1^{t+1} all lie in $\mathcal{X}_\rho^\eta$, so, by Lemma 5.3 and $\alpha^t \leq 1/(1+\zeta)$ (see (52)), (48) holds. Then, using the second inequality in (52) to bound the right-hand side of (48) yields

$$f(x_1^{t+1}) \leq f(x_1^t) - \epsilon_1 \alpha^t \|\hat{r}^t\|^2 - \lambda_m \zeta (\alpha^t \beta^t)^2 + \lambda_m \zeta (\alpha^{t-1} \beta^{t-1})^2 \tag{54}$$

for $t = s, ..., k$. Also, we have that the inequality (42) holds for $h = s$. (For $s = 1$, this follows from (40); for $s > 1$, this follows from α^{s-1} being chosen in Step 1 such that (42) holds for $h = s$.) Summing (54) over all $t \in \{s, ..., k\}$ and then adding to it this inequality, we obtain

$$f(x_1^{k+1}) \leq \eta - \epsilon_1 \sum_{t=1}^{k} \alpha^t \|\hat{r}^t\|^2 - \lambda_m \zeta (\alpha^k \beta^k)^2. \tag{55}$$

The right-hand side of this inequality is less than η, so the claim (53) holds for $t = k$. By induction on k, it follows that (53) holds for $t = s-1, ..., h-1$.

Thus, if (52) holds, then (53) holds for $t = s-1, ..., h-1$, so that (41) holds. In addition, our argument showed that (55) holds for $k = s, ..., h-1$, so that, upon letting $k = h-1$ in (55), we obtain that (42) holds. To see that (43) holds, we use (44) and (24) with $t = s$ to obtain

$$\|r^s - \hat{r}^s\| = \|[x_1^s - \nabla f(x_1^s)]^+ - [x_1^s - g^s]^+\| \leq \|\nabla f(x_1^s) - g^s\| \leq \lambda \alpha^s \beta^s \leq \epsilon_2 \|\hat{r}^s\|,$$

where the first inequality uses the nonexpansive property of $[\cdot]^+$ and the last inequality is due to the last inequality in (52). Thus $\alpha^s, ..., \alpha^{h-1}$, being the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, ...,\}$ such that (41)–(43) hold, are well defined. This completes the induction step and shows that α^t is well defined for all $t = 0, 1, ...$

There are three cases: either (i) $r^s = 0$ for some $s \in T$ or else, since (42) holds for all $h \in T$, (ii) $\liminf_{t\to\infty} f(x_1^t) = -\infty$, or (iii)

$$\sum_{t=1}^{\infty} \alpha^t \|\hat{r}^t\|^2 < \infty. \tag{56}$$

Also, we have from $f(x_1^0) < \eta$ and (39) and (41) (with $h = \min\{t \in T : t > s\}$) for all $s \in T$ that

$$(x_1^t, ..., x_m^t) \in (\mathcal{X}_\rho^\eta)^m, \quad t = 0, 1, ... \tag{57}$$

We claim that, in case (iii), $\{\hat{r}^t\} \to 0$. Since $\{\alpha^t\}$ is monotonically decreasing, either $\liminf_{t\to\infty} \alpha^t > 0$ or $\{\alpha^t\} \downarrow 0$. In the first case, (56) yields $\{\hat{r}^t\} \to 0$. Consider now the second case. We showed earlier that, for each $s \in T$, (41)–(43) hold whenever (52) holds for $t = s, ..., h-1$, where $h = \min\{t \in T : t > s\}$. Then, the choice of $\alpha^s = \cdots = \alpha^{h-1}$ being the largest element of $\{\alpha^{s-1}, \omega\alpha^{s-1}, ...,\}$ such that (41)–(43) hold implies either $\alpha^s = \alpha^{s-1}$ or

$$\frac{\alpha^s}{\omega} > \min_{t=s,...,h-1} \left\{ \min\{\delta_0, 1/(1+\zeta)\}, \frac{(1+\zeta-\epsilon_1)\|\hat{r}^t\|^2}{(1.5\lambda + 3\lambda_m\zeta)(\beta^t)^2}, \frac{\epsilon_2 \|\hat{r}^s\|}{\lambda \beta^s} \right\}. \tag{58}$$

Since (57) holds, then for $t = 0, 1, ...$ we have from (44), (10), (45), and (15) that

$$\|\hat{r}^t\| = \|[x_1^t - g^t]^+ - [x_1^t]^+\| \leq \|g^t\| = \left\|\sum_{i=1}^m \nabla f_i(x_i^t)\right\| \leq \sum_{i=1}^m \beta_i = \beta$$

and from (38), (45), and (15) that

$$\sum_{i=1}^m \|d_i^t\| = \sum_{i=1}^m \left\|\nabla f_i(x_i^t) + \zeta \nabla f_{i\ominus 1}\left(x_{i\ominus 1}^{\lfloor \frac{tm+i-2}{m} \rfloor}\right)\right\| \leq \sum_{i=1}^m (\beta_i + \zeta\beta_{i\ominus 1}) = \beta(1+\zeta),$$

so (10) yields $\beta^t \leq \beta(1+\zeta)$. Since $\{\alpha^s\} \downarrow 0$ so that (58) holds for all s in some subsequence of T, it follows that $\hat{r}^t \to 0$ for t along some subsequence of $\{0, 1, ...\}$. We now argue that $\hat{r}^t \to 0$ for t along the entire sequence $\{0, 1, ...\}$. Suppose this is not the case so there exists $\epsilon > 0$ such that $\|\hat{r}^t\| > \epsilon$ for all t along some subsequence of $\{0, 1, ...\}$. The following argument is a modification of the proof of [12, Prop. 2]. Consider any t such that $\|\hat{r}^t\| \geq \epsilon$. Since $\{\hat{r}^t\}_{t=0,1,...}$ contains a subsequence that tends to zero, there exists a smallest integer $t' > t$ such that $\|\hat{r}^{t'}\| < \epsilon/2$. Then (44) yields

$$\begin{aligned}
\frac{\epsilon}{2} &\leq \|\hat{r}^t\| - \|\hat{r}^{t'}\| \\
&\leq \|\hat{r}^t - \hat{r}^{t'}\| \\
&= \left\| \left[x_1^t - \sum_{i=1}^m \nabla f_i(x_i^t)\right]^+ - x_1^t - \left[x_1^{t'} - \sum_{i=1}^m \nabla f_i(x_i^{t'})\right]^+ + x_1^{t'} \right\| \\
&\leq \left\| \left(x_1^t - \sum_{i=1}^m \nabla f_i(x_i^t)\right) - \left(x_1^{t'} - \sum_{i=1}^m \nabla f_i(x_i^{t'})\right) \right\| + \|x_1^t - x_1^{t'}\| \\
&\leq 2\|x_1^t - x_1^{t'}\| + \left\| \sum_{i=1}^m (\nabla f_i(x_i^t) - \nabla f_i(x_i^{t'})) \right\| \\
&= 2\left\| \sum_{\tau=t}^{t'-1} \alpha^\tau \sum_{i=1}^m d_i^\tau \right\| + \left\| \sum_{\tau=t}^{t'-1} \sum_{i=1}^m \left(\nabla f_i(x_i^{\tau+1}) - \nabla f_i(x_1^{\tau+1}) + \nabla f_i(x_1^{\tau+1}) - \nabla f_i(x_i^\tau)\right) \right\| \\
&\leq 2\sum_{\tau=t}^{t'-1} \alpha^\tau \beta^\tau + \sum_{\tau=t}^{t'-1} \sum_{i=1}^m \lambda_i(\alpha^{\tau+1}\beta^{\tau+1} + 2\alpha^\tau\beta^\tau) \\
&= \sum_{\tau=t}^{t'-1} \left(2\alpha^\tau\beta^\tau + \lambda(\alpha^{\tau+1}\beta^{\tau+1} + 2\alpha^\tau\beta^\tau)\right) \\
&\leq (2+3\lambda)\beta(1+\zeta) \sum_{\tau=t}^{t'-1} \alpha^\tau,
\end{aligned}$$

where the third inequality uses the nonexpansive property of $[\cdot]^+$; the second equality follows from (37); the fifth inequality follows from (10), (57), and Lemma 5.1; the last inequality follows from $\beta^\tau \leq \beta(1+\zeta)$ for all τ and the monotone decreasing property of $\{\alpha^\tau\}_{\tau=0,1,...}$. We also have that $\|\hat{r}^\tau\| \geq \epsilon/2$ for $\tau = t, ..., t'-1$, which together with the above relation yields

$$\sum_{\tau=t}^{t'-1} \alpha^\tau \|\hat{r}^\tau\|^2 \geq \frac{\epsilon^2}{4} \sum_{\tau=t}^{t'-1} \alpha^\tau \geq \frac{\epsilon^2}{4} \frac{\epsilon}{2(2+3\lambda)\beta(1+\zeta)}.$$

Since the number of such t is infinite, this implies $\sum_{\tau=1}^{\infty} \alpha^\tau \|\hat{r}^\tau\|^2 = \infty$, a contradiction of (56).

Since $\{\hat{r}^t\} \to 0$ and (43) holds for all $s \in T$, it follows that $\{r^s\}_{s\in T} \to 0$. □

We may ask whether a convergence rate and stepsize result analogous to Proposition 4.2 holds for the incremental gradient-projection method. This, however, appears unlikely since the proof of Proposition 4.2 requires that $\nabla f_1(x) = \cdots = \nabla f_m(x) = 0$ at a stationary point x of the problem. Such an assumption is reasonable for an unconstrained problem but not for a constrained problem.

6. Implementation issues and numerical experience. To gain some insight into the implementation issues associated with the incremental gradient(-projection) method and its practical performance, we implemented the method to train a single-hidden-layer feedforward neural network, and compared the performance of the method (which, in this case, is effectively on-line backpropagation) with the conjugate gradient method using Polak–Ribiere update and Armijo stepsize rule. In this section we report our findings.

First, we briefly describe the problem of training a single-hidden-layer feedforward neural network (see [12, section 3] for a more detailed discussion). In this problem, we are given a collection of vectors $(I(1), O(1)), ..., (I(m), O(m))$ in $\Re^M \times \Re^L$ ("training examples" of input and desired output), and the goal is to minimize the output error function f given by (1) with $f_i : \Re^{MN+LN+N+L} \mapsto [0, \infty)$, $i = 1, ..., m$, given by

$$f_i(u_1, ..., u_N, v_1, ..., v_N, \omega_1, ..., \omega_N, z) = \left\| \sum_{k=1}^{N} v_k \sigma\left(\langle I(i), u_k \rangle + \omega_k\right) + z - O(i) \right\|^2,$$

where $u_1, ..., u_N \in \Re^M$, $v_1, ..., v_N \in \Re^L$, $\omega_1, ..., \omega_N \in \Re$, $z \in \Re^L$, and with $\sigma : \Re \mapsto \Re$ ("sigmoidal activation function") a user-chosen continuous function satisfying $\sigma(\theta) \to 1$ as $\theta \to \infty$ and $\sigma(\theta) \to 0$ as $\theta \to -\infty$. (In neural network terminology, N is the number of hidden neurons, $u_1, ..., u_N, v_1, ..., v_N$ are the weights on the neural connections, and $\omega_1, ..., \omega_N, z$ are biases at the neurons.) In our testing we used the following standard choice of σ:

$$\sigma(\theta) = 1/(1 + exp(-\theta/10))$$

and set N according to the specific training examples, as is done in practice. We have two specific test problems: For our first test problem, that of computing the parity of $\{0, 1\}$-vectors [10, p. 131], we have $m = 5$, $M = 4$, $L = 1$, and

$$I(1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, O(1) = 0, \; I(2) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, O(2) = 1, \; I(3) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, O(3) = 0, \; \cdots,$$

$$I(5) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, O(5) = 0,$$

and we set $N = 1$. (Here the desired output is 0 when the input has an even number of 1's and, otherwise, the desired output is 1.) For our second test problem, that of

recognizing the characters of 0, 1, 2, 3, 4, we have $m = 5$, $M = 15$, $L = 3$, and

$$I(1) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \; O(1) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \cdots, \qquad I(5) = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \; O(5) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and we set $N = 3$. (Here the "1"s in $I(1)$, written as a 5×3 matrix, form the character 0 and similarly for $I(2), ..., I(5)$. The desired output $O(i)$ is the binary representation of $i - 1$, for $i = 1, ..., 5$.) We also experimented with a version of the problem in which the desired output was changed to $L = 1$ and $O(1) = 0, ..., O(5) = 4$. However, though all methods converged faster on the problem, the trained neural network was much less accurate in recognizing corrupted input. Two key features of each function f_i are that, (i) it has multiple local minima, and (ii) evaluating ∇f_i requires roughly the same work as evaluating f_i.

Next, we describe the three methods we implemented. The first method, referred to as ALG1, is the incremental gradient method of section 2. For this method, we used the standard choice of $\epsilon_0 = 1$, $\omega = .5$, $T = \{1, 11, 21, 31, ...\}$ and, after some experimentation, found the choice of $\zeta = .8$ to work best (much better than the memoryless choice of $\zeta = 0$). To avoid stepsizes becoming too small, it is desirable to choose η, ϵ_3 large and ϵ_1, ϵ_2 small and, for our implementation, we chose $\eta = 1.5 f(x_1^0) + 100$, $\epsilon_3 = 1000$, $\epsilon_1 = \epsilon_2 = .00001$ and estimated $\lambda_1 + \cdots + \lambda_m$ and ρ by 1 and ∞, respectively. The second method, referred to as ALG2, is the incremental gradient-projection method of section 5 (with $\mathcal{X} = \Re^n$). For this method, we used the standard choice of $\epsilon_0 = 1$, $\omega = .5$, $T = \{1, 11, 21, 31, ...\}$ and, after some experimentation, found the choice of $\zeta = .1$ to work best. Analogous to ALG1, we chose $\eta = 1.5 f(x_1^0) + 100$, $\epsilon_1 = .00001$, $\epsilon_2 = 1000$, and estimated λ_m and ρ by 1 and ∞, respectively. The third method, referred to as ALG3, is the conjugate gradient method using the efficient Polak–Ribiere update and the Armijo stepsize rule [1, pp. 20, 57]. More precisely, the method generates, for any given $x^0 \in \Re^n$, a sequence $x^0, x^1, ...$ according to

$$x^{t+1} := x^t - \alpha^t d^t,$$

where

$$d^t := \begin{cases} \nabla f(x^t) + \beta^t d^{t-1} & \text{if } t > 0, \\ \nabla f(x^t) & \text{if } t = 0, \end{cases}$$

$$\beta^t := \langle \nabla f(x^t), \nabla f(x^t) - \nabla f(x^{t-1}) \rangle / \|\nabla f(x^{t-1})\|^2,$$

and α^t is the largest element of $\{\epsilon_0, \epsilon_0\omega, \epsilon_0(\omega)^2, ...\}$ for which x^{t+1} given above satisfies

$$f(x^{t+1}) \leq f(x^t) - \sigma\alpha^t \langle \nabla f(x^t), d^t \rangle.$$

Here $\epsilon_0 > 0, \omega \in (0, 1), \sigma \in (0, .5)$ are user-chosen parameters. In our implementation, we used the standard choice of $\epsilon_0 = 1, \omega = .5, \sigma = .1$, and, to ensure convergence, incorporated a steepest descent restart (i.e., replace d^t by $\nabla f(x^t)$ whenever $\langle \nabla f(x^t), d^t \rangle < .00001 \|\nabla f(x^t)\| \|d^t\|$). For all three methods, each component of the starting point (i.e., x_1^0 for ALG1, ALG2, and x^0 for ALG3) was randomly generated according to the uniform distribution on the interval $[0, 1]$ and the termination criterion was $f(x) \leq 10^{-8}$. All methods were coded in Matlab Version 4.2a and were run

TABLE 1
Performance of ALG1, ALG2, ALG3 *on the two test problems.*

	ALG1		**ALG2**		**ALG3**	
Problem	**ngrad.**[1]	**nfunc**[2]	**ngrad**[1]	**nfunc**[2]	**ngrad**[1]	**nfunc**[2]
Parity	222.3	30.2	602.3	61.3	650.2	1775.0
Character Recognition	185.6	19.6	629.0	64.0	227.0	713.3

[1] ngrad denotes the number of times that $\nabla f_1, ..., \nabla f_m$ have been evaluated.
[2] nfunc denotes the number of times that $f_1, ..., f_m$ have been evaluated.

on a Decstation 5000. Table 1 gives the number of gradient and function evaluations for the three methods, averaged over three runs (all with a standard deviation of less than 20 percent).

From Table 1, it can be seen that ALG1 requires fewer gradient evaluations and function evaluations (which are the most expensive operations) than either ALG2 or ALG3. Since gradient evaluation requires roughly equal work as function evaluation so that the total work is roughly equal to the sum of gradient and function evaluations, we see that ALG1 requires less than one-third the total work of either ALG2 or ALG3, while ALG3 requires the most total work. Thus, for our test problems at least, we can draw the following conclusions: (i) the incremental gradient method using an unlimited-memory momentum term is more efficient than the conjugate gradient method using Polak–Ribiere update and Armijo stepsize rule; (ii) the incremental gradient method using an unlimited-memory momentum term is more efficient than that using a one-memory momentum term (which in turn is more efficient than that using no momentum term at all). We note that the stepsize rule is also crucial to the efficiency of the incremental gradient method. When we took ALG1 and replaced its stepsize rule by the well-known (nonadapative) stepsize rule of

$$\alpha^t = c/t$$

(which produces stepsizes that are square summable but not summable), the convergence became agonizingly slow, regardless of the choice of the constant $c > 0$. In contrast, the stepsizes in both ALG1 and ALG2 remained at the value .5 after an initial decrease (the large stepsizes, as well as the presence of the momentum term, appear to be key to the good performance of ALG1), while the stepsizes in ALG3 varied between .125 and 1. On the other hand, we caution that these results are for some small test problems only and much more extensive testing is needed to determine the efficiency of the incremental gradient(-projection) method in general.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[2] D. P. BERTSEKAS, *Incremental least squares methods and the extended Kalman filter*, SIAM J. Optim., 6 (1996), pp. 807–822.
[3] D. P. BERTSEKAS, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.
[4] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[5] W. C. DAVIDON, *New least-square algorithms*, J. Optim. Theory Appl., 18 (1976), pp. 187–197.
[6] T. DENOEUX AND R. LENGELLÉ, *Initializing back propagation networks with prototypes*, Neural Networks, 6 (1993), pp. 351–363.

[7] E. M. Gafni and D. P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
[8] A. A. Gaivoronski, *Convergence properties of back-propagation for neural nets via theory of stochastic gradient methods, Part* I, Optim. Methods Software, 4 (1994), pp. 117–134.
[9] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Software, 4 (1994), pp. 135–150.
[10] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison–Wesley, Redwood City, CA, 1991.
[11] Y. Le Cun, *Une procedure d'apprentissage pour reseau a seuil assymetrique*, in Proc. Cognitiva '85, Paris, France, pp. 599–604.
[12] Z.Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Software, 4 (1994), pp. 85–101.
[13] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Software, 4 (1994), pp. 103–116.
[14] M. F. Møller, *A scaled conjugate gradient algorithm for fast supervised learning*, Neural Networks, 6 (1993), pp. 525–533.
[15] T. N. Pappas, *Solution of Nonlinear Equations by Davidon's Least Squares Method*, M.Sc. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1982.
[16] D. B. Parker, *Learning-Logic*, Center for Computational Research in Economics and Management Science Report no. TR-47, Massachusetts Institute of Technology, Cambridge, MA, 1985.
[17] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*, in Parallel Distributed Processing–Explorations in the Microstructure of Cognition, Rumelhart and McClelland, eds., MIT Press, Cambridge, MA, 1986, pp. 318–362.
[19] G. Tesauro, Y. He, and S. Ahmad, *Asymptotic convergence of back propagation*, Neural Comput., 1 (1989), pp. 382–391.
[20] P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, Ph.D. thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA, 1974.
[21] P. J. Werbos, *Backpropagation through time: What it does and how to do it*, in Proc. IEEE, 78 (1990), pp. 1550–1560.
[22] H. White, *Learning in artificial neural networks: A statistical perspective*, Neural Comput., 1 (1989), pp. 425–464.
[23] H. White, *Some asymptotic results for learning in single hidden-layer feedforward network models*, J. Amer. Statist. Assoc., 84 (1989), pp. 1003–1013.