Proceedings of the 38th
Conference on Decision & Control
Phoenix, Arizona USA • December 1999

TuP02 16:20

INCREMENTAL SUBGRADIENT METHODS FOR NONDIFFERENTIABLE OPTIMIZATION

Angelia Geary Dimitri P. Bertsekas
angeg@lids.mit.edu bertsekas@lids.mit.edu

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

ABSTRACT

We propose a new class of subgradient methods for minimizing a convex function that consists of the sum of a large number of component functions. This type of minimization arises in a dual context from Lagrangian relaxation of the coupling constraints of large scale separable problems. The idea is to perform the subgradient iteration incrementally, by sequentially taking steps along the subgradients of the component functions, with intermediate adjustment of the variables after processing each component function. This incremental approach has been very successful in solving large differentiable least squares problems, such as those arising in the training of neural networks, and it has resulted in a much better practical rate of convergence than the steepest descent method.

In this paper, we establish the convergence properties of a number of variants of incremental subgradient methods, including some that are stochastic. Based on the analysis and computational experiments, the methods appear very promising and effective for important classes of large problems.

1 Introduction

Throughout this paper, we focus on the problem

$$\text{minimize} \quad f(x) = \sum_{i=1}^{m} f_i(x)$$

$$\text{subject to} \quad x \in X, \tag{1}$$

where $f_i : \Re^n \to \Re$ are convex functions and X is a closed and convex subset of $\Re^n$. We are primarily interested in the case where f is nondifferentiable. A special case of particular interest to us is when f is the dual function of a primal separable combinatorial problem of the form

$$\text{maximize} \quad \sum_{i=1}^{m} c_i' y_i$$

$$\text{subject to} \quad y_i \in Y_i, \quad i = 1, \ldots, m,$$

$$\sum_{i=1}^{m} A_i y_i \geq b,$$

where Y_i is a given finite subset of $\Re^p$, c_i are given vectors in $\Re^p$, A_i are given $n \times p$ matrices and b is a given vector in $\Re^n$. By viewing x as a Lagrange multiplier vector for the coupling constraint $\sum_{i=1}^{m} A_i y_i \geq b$, we obtain a dual problem of the form (1), where

$$f_i(x) = \beta_i' x + \min_{y_i \in Y_i} (c_i - A_i' x)' y_i, \tag{2}$$

and β_i are vectors in $\Re^n$ such that

$$\beta_1 + \cdots + \beta_m = b,$$

and prime denotes transposition. It is well-known that solving dual problems of the type above, possibly in a branch-and-bound context, is one of the most important and challenging algorithmic areas of optimization.

A principal method for solving problem (1) is the subgradient method

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha_k \sum_{i=1}^{m} d_{i,k} \right] \tag{3}$$

where $d_{i,k}$ is a subgradient of f_i at x_k, α_k is a positive stepsize, and $\mathcal{P}_X$ denotes projection on X. There is an extensive theory for this method (see Dem'yanov and Vasil'ev [DeV85], Shor [Sho85], Minoux [Min86], Polyak [Pol87], Hiriart-Urruty and Lemaréchal [HiL93], Bertsekas [Ber99]). In many important applications, the set X is simple enough so

0-7803-5250-5/99/$10.00 © 1999 IEEE

that the projection can be easily implemented. In particular, for the special case of the dual problem (1), (2), the set X is the positive orthant $\{x \in \Re^n \mid x \geq 0\}$ and projecting on X is not expensive.

The incremental subgradient method that we propose is similar to the standard subgradient method (3). The main difference is that at each iteration, x is changed incrementally, through a sequence of m steps. Each step is a subgradient iteration for a single component function f_i, and there is one step per component function. Thus, an iteration can be viewed as a cycle of m subiterations. If x_k is the vector obtained after k cycles, the vector x_{k+1} obtained after one more cycle is

$$x_{k+1} = \psi_{m,k}, \tag{4}$$

where $\psi_{m,k}$ is obtained after the m steps

$$\psi_{i,k} = \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k g_{i,k}\right], \quad i = 1, \ldots, m, \tag{5}$$

starting with

$$\psi_{0,k} = x_k, \tag{6}$$

where $g_{i,k}$ is a subgradient of f_i at the point $\psi_{i-1,k}$.

Incremental gradient methods for *differentiable* unconstrained problems have a long tradition, most notably in the training of neural networks, where they are known as *backpropagation methods*. They are related to the Widrow-Hoff algorithm [WiH60] and to stochastic gradient/stochastic approximation methods, and they are supported by several recent convergence analyses (Luo [Luo91], Grippo [Gri94], Gaivoronski [Gai94], Luo and Tseng [LuT94], Mangasarian and Solodov [MaS94], Bertsekas and Tsitsiklis [BeT96], Bertsekas [Ber97], Bertsekas and Tsitsiklis [BeT99]). It has been experimentally observed that incremental gradient methods often converge much faster than the steepest descent method when far from the eventual limit. However, near convergence, they typically converge slowly because they require a diminishing stepsize [e.g. $\alpha_k = O(1/k)$] for convergence. If α_k is instead taken to be a small enough constant, "convergence" to a limit cycle occurs, as first shown by Luo [Luo91].

We will propose a variety of stepsize rules and we will give a number of convergence results. Just like the differentiable case, a diminishing stepsize is essential for the convergence of the incremental subgradient method. To understand the reason it is helpful to view the incremental subgradient method as an approximate subgradient method (or a subgradient method with errors). In particular, since we have $\|\psi_{i,k} - x_k\| = O(\alpha_k)$, it can be seen that if the iterates x_k are bounded, then the direction $\sum_{i=1}^{m} g_{i,k}$ used by the incremental subgradient method is an ϵ_k-subgradient of f at x_k, where $\epsilon_k > 0$ is proportional to α_k. Hence ϵ_k is diminishing to zero only if α_k tends to 0. If $\alpha_k \to 0$ and some additional conditions, such as $\sum_{k=0}^{\infty} \alpha_k = \infty$ hold, then the incremental method exhibits convergence behavior similar to methods that use ϵ-subgradients (see Dem'yanov and Vasil'ev [DeV85], Polyak [Pol87], p. 144, Correa and Lemaréchal [CoL93], Hiriart-Urruty and Lemaréchal [HiL93], Bertsekas [Ber99]). Nevertheless, our results include stepsize rules where the condition $\sum_{k=0}^{\infty} \alpha_k = \infty$ is not explicitly imposed.

The paper is organized as follows. In the next section, we present the convergence results of the incremental subgradient method for different stepsize rules. In Section 3, we establish the convergence properties of randomized versions of the incremental subgradient method.

The proofs of the results given here can be found in [GeB99]. Furthermore, under additional assumptions on f, we have obtained estimates of the convergence rates, which are also presented in [GeB99].

2 An Incremental Subgradient Method

Throughout this paper, we use the following notation:

$$f^* = \inf_{x \in X} f(x), \qquad X^* = \{x \in X \mid f(x) = f^*\},$$

$$f_i^* = \inf_{x \in X} f_i(x), \qquad 1 \leq i \leq m.$$

We assume that the following holds.

Assumption 2.1:

(a) For each i the value f_i^* is finite and for some i_0 the level sets

$$L_0(c) = \{x \in X \mid f_{i_0}(x) \leq c\}$$

of the function f_{i_0} are bounded.

(b) The set of subgradients $\{g_{i,k}\}$ used by the incremental subgradient method (4)-(6) is bounded, i.e., there exists a positive constant C_0 such that

$$\|g_{i,k}\| \leq C_0, \quad i = 1, \ldots, m, \quad k = 0, 1, \ldots.$$

In many important applications, the set X is compact so that Assumption 2.1 is satisfied [the set $\cup_{x \in X} \partial f_i(x)$ is compact if X is compact]. Also, the condition (b) is satisfied if each f_i is polyhedral (i.e., f_i is the pointwise minimum of a finite number of affine functions). In particular, condition (b) holds for the dual problem (1), (2), where for all i and x the set of subgradients $\partial f_i(x)$ is the convex hull of a finite number of points.

We first consider the behavior of the incremental subgradient method with a constant stepsize.

Proposition 2.1: *Let Assumption 2.1 hold. Furthermore, suppose that the stepsize in Eqs. (4)-(6) is fixed, i.e., $\alpha_k \equiv \alpha$ for some positive constant α. Then we have*

$$\liminf_{k\to\infty} f(x_k) \le f^* + \frac{\alpha C^2}{2},$$

where

$$C = \sum_{i=1}^{m} C_i,$$

$$C_i = \max\Big\{C_0,\ \max_{k\ge 0}\{\|g\| \mid g \in \partial f_i(x_k)\}\Big\}.$$

It can be seen that if the stepsize α_k is bounded, then the sequence $\{x_k\}$ generated by the incremental subgradient method (4)-(6) is also bounded. Hence the constants C_i above are finite.

The next proposition parallels a well-known convergence result for the ordinary subgradient method (see Shor [Sho85], p. 25).

Proposition 2.2: *Let Assumption 2.1 hold and assume that the stepsize α_k is such that*

$$\alpha_k > 0, \qquad \lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method (4)-(6), we have

$$\lim_{k\to\infty} dist(x_k, X^*) = 0, \qquad \lim_{k\to\infty} f(x_k) = f^*,$$

where $dist(y, X^*)$ denotes the Euclidean distance from a point y to the set X^*.

Proposition 2.2 gives sufficient conditions to ensure that all limit points of $\{x_k\}$ are optimal solutions. However, additional restrictions on the stepsize are needed to have the convergence of the iterates. The following proposition states such conditions.

Proposition 2.3: *Let Assumption 2.1 hold, and assume that the stepsize α_k is such that*

$$\alpha_k > 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then the sequence $\{x_k\}$ converges to a point $\overline{x} \in X^*$.

All the results presented so far rely on Assumption 2.1. Therefore the preceding results do not apply to the simple case where each f_i is polyhedral but none of the f_i has bounded level sets. In the next proposition, under assumption that each f_i is polyhedral, we give the convergence result that is somewhat weaker than the preceding results, but is valid even if f^* is not finite. Moreover, the result is applicable to the case where all components f_i are polyhedral and for some of them we have $f_i^* = -\infty$. In particular, the dual problem (1), (2) may have such components.

Proposition 2.4: *Assume that each f_i is polyhedral. Let the stepsize α_k be such that*

$$\alpha_k > 0, \qquad \lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method (4)-(6), we have

$$\liminf_{k\to\infty} f(x_k) = f^*.$$

The preceding results apply to the constant and the diminishing stepsize choices. An interesting alternative for the ordinary subgradient method is the dynamic stepsize rule

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{\|g_k\|^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2,$$

with g_k being a subgradient of f at x_k, suggested by B. T. Polyak in [Pol69], (see also discussions in Shor [Sho85], Brannlund [Bra93], and Bertsekas [Ber99]). For the incremental method, we propose a variant of this stepsize that has the form

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{C^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2. \quad (7)$$

where

$$C = \sum_{i=1}^{m} C_i, \quad (8)$$

$$C_i \ge \max_{k\ge 0}\{\|g\| \mid g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k})\}. \quad (9)$$

For this choice of stepsize we have to be able to calculate the upper bounds C_i, which can be done, for example, when the components f_i are polyhedral.

For the rest of this section, we assume that the following holds.

Assumption 2.2: *The upper bounds*

$$C_i \ge \max_{k\ge 0}\{\|g\| \mid g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k})\},$$

for $i = 1, \ldots, m$ are finite.

Note that if each f_i is polyhedral or if X is compact, then Assumption 2.2 is satisfied. We first give our results assuming that f^* is known. We later modify the stepsize, so that f^* can be replaced by a dynamically updated estimate.

Proposition 2.5: *Let Assumption 2.2 hold. Also, assume that the optimal cost f^* is finite and the optimal set X^* is nonempty. Then the sequence $\{x_k\}$ obtained by the incremental subgradient method (4)-(6) with the dynamic stepsize (7)-(9) converges to a solution $\overline{x} \in X^*$.*

In most practical problems the value f^* is not known. To address this, we may modify the dynamic stepsize (7) by using an estimate for f^* in place of its exact value. This leads to incremental subgradient method (4)–(6) with modified dynamic stepsize

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{C^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2, \quad (10)$$

with C defined by Eqs. (8) and (9), and $f_k^{\rm lev}$ being a target level estimate of f^*.

We discuss two procedures for updating the target values $f_k^{\rm lev}$. In both procedures $f_k^{\rm lev}$ is equal to the best function value $\min_{0 \le j \le k} f(x_j)$ achieved up to the k-th iteration minus a positive amount δ which is adjusted based on the algorithm's progress. The first adjustment procedure is simple but is only guaranteed to yield a δ-optimal objective function value with δ positive and arbitrarily small, (unless $f^* = -\infty$ in which case the procedure yields the optimal function value). The second adjustment procedure for $f_k^{\rm lev}$ is more complex but is guaranteed to yield the optimal value f^* in the limit. This procedure is based on the ideas and algorithms of Brannlund [Bra93], and Goffin and Kiwiel [GoK99].

In the first adjustment procedure $f_k^{\rm lev}$ is

$$f_k^{\rm lev} = \min_{0 \le j \le k} f(x_j) - \delta_k, \qquad (11)$$

where δ_k is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k & \text{if } f(x_{k+1}) < f_k^{\rm lev}, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) \ge f_k^{\rm lev}, \end{cases} \qquad (12)$$

where δ, β, and ρ are fixed positive constants with $\beta < 1$ and $\rho \ge 1$. Thus in this procedure we essentially "aspire" to reach a target level that is smaller by δ_k over the best value achieved thus far. Whenever the target level is achieved, we increase δ_k or we keep it at the same value depending on the choice of ρ. If the target level is not attained at a given iteration, δ_k is reduced up to a threshold δ. This threshold guarantees that the stepsize α_k of Eq. (10) is bounded away from zero, since we have

$$\alpha_k \ge \underline{\gamma}\,\frac{\delta}{C^2}.$$

The effect is that the method behaves similar to the one with a constant stepsize (cf. Proposition 2.1). In particular, we have the following result.

Proposition 2.6: *Let Assumption 2.2 hold, and let $\{x_k\}$ be a sequence generated by the incremental target level method [cf. Eqs. (4)-(6) and (10)] with the adjustment procedure (11), (12).*

(a) If the optimal value f^* is not finite, then we have

$$\inf_{k \ge 0} f(x_k) = -\infty = f^*.$$

(b) If f^* is finite, then

$$\inf_{k \ge 0} f(x_k) \le f^* + \delta.$$

The adjustment procedure for $f_k^{\rm lev}$ given by Eqs. (11) and (12) can also be used for the ordinary subgradient method. In particular, suppose we want to minimize a convex function $h : \Re^n \to \Re$ over a convex and closed subset X of $\Re^n$. Consider the following subgradient method

$$x_{k+1} = \mathcal{P}_X[x_k - \alpha_k g_k], \qquad \forall\, k \ge 0, \qquad (13)$$

where $x_0 \in X$ is a given initial point, g_k is a subgradient of h at point x_k, and the stepsize α_k is

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{\|g_k\|^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2, \quad (14)$$

with the target level $f_k^{\rm lev}$ defined by Eqs. (11), (12). It can be seen that Proposition 2.6 holds for $\{x_k\}$ generated by the subgradient method (13), (14).

We now consider another procedure for adjusting $f_k^{\rm lev}$, which guarantees that $f_k^{\rm lev} \to f^*$, and convergence of the associated method to the optimum. In this procedure we reduce δ_k whenever the method "travels" for a long distance without reaching the corresponding target level.

Incremental Target Level Algorithm

Step 0 (*Initialization*) Select x_0, $\delta_0 > 0$, and $B > 0$. Set $\sigma_0 = 0$, $f_{-1}^{\rm rec} = \infty$. Set $k = 0$, $l = 0$, and $k(l) = 0$ [$k(l)$ will denote the iteration number when the l-th update of $f_k^{\rm lev}$ occurs].

Step 1 (*Function evaluation*) Calculate $f(x_k)$. If $f(x_k) < f_{k-1}^{\rm rec}$, then set $f_k^{\rm rec} = f(x_k)$. Otherwise set $f_k^{\rm rec} = f_{k-1}^{\rm rec}$ [so that $f_k^{\rm rec}$ keeps the record of the smallest value attained by the iterates that are generated so far, i.e. $f_k^{\rm rec} = \min_{0 \le j \le k} f(x_j)$].

Step 2 (*Sufficient descent*) If $f(x_k) \le f_{k(l)}^{\rm rec} - \frac{\delta_l}{2}$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \delta_l$, increase l by 1, and go to Step 4.

Step 3 (*Oscillation detection*) If $\sigma_k > B$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \frac{\delta_l}{2}$, and increase l by 1.

Step 4 (*Iterate update*) Set $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$. Select $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and calculate x_{k+1} via Eqs. (4)–(6), (10).

Step 5 (*Path length update*) Set $\sigma_{k+1} = \sigma_k + C\alpha_k$. Increase k by 1 and go to Step 1.

In this procedure, the target level is updated only if sufficient descent or oscillation is detected (cf. Step 2 or Step 3). It can be seen that σ_k is an upper bound on the length of the path traveled by the iterates $x_{k(l)}, \ldots, x_k$. Whenever σ_k exceeds the prescribed upper bound B on the path length, the target level is decreased. Here is the result regarding the convergence of the incremental subgradient target level algorithm.

Proposition 2.7: *If Assumption 2.2 holds, then for the incremental target level algorithm we have*

$$\lim_{k\to\infty} f_k^{\text{rec}} = \inf_{k\geq 0} f(x_k) = f^*.$$

3 An Incremental Subgradient Method with Randomization

In the preceding convergence analysis, all the results given are valid regardless of the order in which the component functions f_i are processed, as long as each component is taken into account exactly once within a cycle. However, the order of processing the components can significantly affect the convergence rate of the method. Unfortunately, to determine the order which results in a favorable convergence rate may be very difficult in practice. A popular technique for incremental gradient methods (for differentiable components f_i) is to reshuffle randomly the order of the functions f_i at the beginning of each cycle. A variation of this method is to pick randomly a function f_i at each iteration rather than to pick each f_i exactly once in every cycle according to a randomized order. This variation can be viewed as a gradient method with random errors, as shown in Bertsekas and Tsitsiklis [BeT96], p. 143. Similarly, the corresponding incremental subgradient method at each step picks randomly a function f_i to be processed next. The analysis of the method can then be performed by following known lines of analysis for stochastic subgradient methods (see e.g., Ermoliev [Erm69], Polyak [Pol87], p. 159, Ermoliev and Wets [ErW88]).

The formal description of the randomized incremental subgradient method is as follows

$$x_k = \psi_{0,k}, \tag{15}$$

$$\psi_{i,k} = \mathcal{P}_X\left[\psi_{i-1,k} - \alpha_k g(\omega_{i,k}, \psi_{i-1,k})\right], \tag{16}$$

$$x_{k+1} = \psi_{m,k}, \tag{17}$$

where $i = 1, \ldots, m$, $x_0 \in X$ is a given point, $\omega_{i,k}$ is a random variable taking equiprobable values from the set $\{1, \ldots, m\}$, and $g(\omega_{i,k}, \psi_{i-1,k})$ is a subgradient of $f_{\omega_{i,k}}$ at $\psi_{i-1,k}$. This simply means that if the random variable $\omega_{i,k}$ takes a value j, then the vector $g(\omega_{i,k}, \psi_{i-1,k})$ is a subgradient of f_j at $\psi_{i-1,k}$.

Proposition 3.1: *Suppose that:*

(a) The optimal cost f^* is finite and the optimal set X^* is nonempty.

(b) The sequence $\{\omega_{i,k}\}$ is a sequence of independent random variables each of which is uniformly distributed over the set $\{1, \ldots, m\}$. Furthermore, the sequence $\{\omega_{i,k}\}$ is independent of the sequence $\{\psi_{i,k}\}$.

(c) The subgradients $g(\omega_{i,k}, \psi_{i-1,k})$ in Eq. (16) are bounded, i.e., there exists a positive constant C_0 such that for all i, k

$$||g(\omega_{i,k}, \psi_{i-1,k})|| \leq C_0,$$

with probability 1.

(d) The stepsize α_k satisfies

$$\alpha_k > 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then the sequence $\{x_k\}$ generated by the randomized incremental subgradient method (15)–(17) converges to a solution $\overline{x} \in X^*$ with probability 1.

The randomized incremental method (15)–(17) can also be used with the dynamic stepsize. Conditions for convergence of this version of the method are given in the following proposition.

Proposition 3.2: *Let the assumptions (a) and (b) of Proposition 3.1 hold. Furthermore, assume that the stepsize α_k is given by*

$$\alpha_k = \gamma_k \frac{m}{2m-1} \frac{f(x_k) - f^*}{C^2}, \tag{18}$$

where $0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2$ and

$$C = \sum_{j=1}^{m} C_j, \qquad C_j \geq \max_{i,k}\{||g|| \mid g \in \partial f_j(\psi_{i,k})\}.$$

Then the sequence $\{x_k\}$ generated by the randomized incremental subgradient method (15)–(17) converges to an optimal solution $\overline{x} \in X^*$ with probability 1.

In Proposition 3.2 we assumed that the optimal value f^* is known. If f^* is unknown, then we can use a dynamically updated target value f_k^{lev} instead of f^* in Eq. (18). The target level f_k^{lev} can be selected according to one of the two adjustment procedures that we discussed in Section 2.

4 Conclusions

The choice of the stepsize α_k plays an important role in the performance of incremental subgradient methods. If the stepsize is constant, the incremental method typically exhibits oscillatory behavior, with a size of oscillation (the diameter of the limit cycle) that obeys the worst case estimate of Proposition 2.1. Nonetheless, the variant of the method with a constant stepsize is important for theoretical and practical reasons. For example, it is common to use a stepsize α_k which is diminishing only up to the point where α_k crosses a certain threshold, and then to keep α_k constant at that threshold.

The performance of the incremental methods is also affected by the *order* in which the components f_i are processed within a cycle. Our computational experiments and our analysis of some examples show that the randomized incremental subgradient method has *qualitatively superior* convergence characteristics than the nonrandomized variant (unless the order of processing the components f_i is favorably chosen).

We observed several common characteristics of the incremental methods, which tend to manifest themselves in some generality:

(a) When far from the solution, the method can make much faster progress than the nonincremental method, particularly if the number of component functions m is large. The rate of progress also depends on the stepsize.

(b) When close to the solution, the method oscillates and the size of the oscillation is proportional to the stepsize. Thus there is a tradeoff between rapid initial convergence (large stepsize) and size of asymptotic oscillation (small stepsize). With a diminishing stepsize the method is capable of attaining convergence (no asymptotic oscillation).

(c) The size of the oscillation depends also on the order in which the component functions f_i are processed within a cycle. The precise effect of the processing order is not clearly understood at present, but it is interesting and substantial. In particular, it appears that for nondifferentiable problems, a randomized order is often superior to a fixed deterministic order.

5 References

[Ber97] Bertsekas, D. P., "A New Class of Incremental Gradient Methods for Least Squares Problems," SIAM J. on Optimization, Vol. 7, 1997, pp. 913–926.

[Ber99] Bertsekas, D. P., Nonlinear Programming, (2nd edition), Athena Scientific, Belmont, MA, 1999.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, MA., 1996.

[BeT99] Bertsekas, D. P., and Tsitsiklis, J. N., "Gradient Convergence in Gradient Methods," to appear in SIAM J. on Optimization, 1999.

[Bra93] Brannlund, U., "On Relaxation Methods for Nonsmooth Convex Optimization," Doctoral Thesis, Royal Institute of Technology, Sweden, 1993.

[CoL93] Correa, R., and Lemaréchal, C., "Convergence of Some Algorithms for Convex Minimization," Math. Programming, Vol. 62, 1993, pp. 261–275.

[DeV85] Dem'yanov, V. F., and Vasil'ev , L. V., Nondifferentiable Optimization, Optimization Software, New York, 1985.

[Erm69] Ermoliev, Yu. M., "On the Stochastic Quasi-gradient Method and Stochastic Quasi-Feyer Sequences," Kibernetika, 1969, No. 2, pp. 73–83.

[ErW88] Ermoliev, Yu. M., and Wets, R. J-B (Eds.), Numerical Techniques for Stochastic Otimization, 1988, IIASA, Springer-Verlag.

[Gai94] Gaivoronski, A. A., "Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks," Optimization Methods and Software, Vol. 4, 1994, pp. 117–134.

[GeB99] Geary, A., and Bertsekas, D. P., "Incremental Subgradient Methods for Nondifferentiable Optimization," Lab. for Info. and Decision Systems Report LIDS-P-2460, MIT, Cambridge, MA, 1999; to appear in SIAM J. on Optimization.

[GoK99] Goffin, J., and Kiwiel, K., "Convergence of a Simple Subgradient Level Method," Math. Programming, Vol. 85, 1999, pp. 207–211.

[Gri94] Grippo, L., "A Class of Unconstrained Minimization Methods for Neural Network Training," Optim. Methods and Software, Vol. 4, 1994, pp. 135–150.

[HiL93] Hiriart-Urruty, J.-B., and Lemaréchal, C., Convex Analysis and Minimization Algorithms, Vols. I and II, Springer-Verlag, Berlin and N.Y., 1993.

[Luo91] Luo, Z. Q., "On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks," Neural Computation, Vol. 3, 1991, pp. 226–245.

[LuT94] Luo, Z. Q., and Tseng, P., "Analysis of an Approximate Gradient Projection Method with Applications to the Backpropagation Algorithm," Optim. Methods and Software, Vol. 4, 1994, pp. 85–101.

[MaS94] Mangasarian, O. L., and Solodov, M. V., "Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization," Optim. Methods and Software, Vol. 4, 1994, pp. 103–116.

[Min86] Minoux, M., Mathematical Programming: Theory and Algorithms, J. Wiley, N.Y., 1986.

[Pol69] Polyak, B. T, "Minimization of Unsmooth Functionals," Z. Vychisl. Mat. i Mat. Fiz., Vol. 9, No. 3, 1969, pp. 509–521.

[Pol87] Polyak, B. T, Introduction to Optimization, Optimization Software Inc., N.Y., 1987.

[Sho85] Shor, N. Z., Minimization Methods for Nondifferentiable Functions, Springer, Berlin, 1985.

[WiH60] Widrow, B., and Hoff, M. E., "Adaptive Switching Circuits," Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, part 4, 1960, pp. 96–104.

Research supported by NSF under Grant ACI-9873339.