Sept. 1999, Revised Feb., 2001 **LIDS-P-2460**

To Appear in SIAM J. on Optimization

INCREMENTAL SUBGRADIENT METHODS[1] FOR NONDIFFERENTIABLE OPTIMIZATION

by

Angelia Nedić and Dimitri P. Bertsekas [2]

Abstract

We consider a 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. A particularly interesting discovery is that by randomizing the order of selection of component functions for iteration, the convergence rate is substantially improved.

[1] Research supported by NSF under Grant ACI-9873339.

[2] Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, Mass., 02139.

1. INTRODUCTION

Throughout this paper, we focus on the problem

$$\begin{aligned} &\text{minimize} \quad f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in X, \end{aligned} \tag{1.1}$$

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

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{m} c_i' y_i \\ &\text{subject to} \quad y_i \in Y_i, \quad i = 1, \ldots, m, \qquad \sum_{i=1}^{m} A_i y_i \geq b, \end{aligned}$$

where prime denotes transposition, c_i are given vectors in $\Re^p$, Y_i is a given finite subset of $\Re^p$, A_i are given $n \times p$ matrices, and b is a given vector in $\Re^n$. Then, 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 (1.1), where

$$f_i(x) = \max_{y_i \in Y_i} (c_i + A_i' x)' y_i - \beta_i' x, \qquad i = 1, \ldots, m, \tag{1.2}$$

β_i are vectors in $\Re^n$ such that $\beta_1 + \cdots + \beta_m = b$, and X is the positive orthant $\{x \in \Re^n \mid x \geq 0\}$. 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 (1.1) is the subgradient method

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha_k \sum_{i=1}^{m} d_{i,k} \right], \tag{1.3}$$

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 the set X. There is an extensive theory for this method (see e.g., the textbooks by 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 that the projection can be easily implemented. In particular, for the special case of the dual problem (1.1), (1.2), the set X is the positive orthant and projecting on X is not expensive.

The incremental subgradient method is similar to the standard subgradient method (1.3). 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

$$x_{k+1} = \psi_{m,k}, \tag{1.4}$$

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

$$\psi_{i,k} = \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k g_{i,k}\right], \qquad g_{i,k} \in \partial f_i(\psi_{i-1,k}), \qquad i = 1, \ldots, m, \tag{1.5}$$

starting with

$$\psi_{0,k} = x_k, \tag{1.6}$$

where $\partial f_i(\psi_{i-1,k})$ denotes the subdifferential (set of all subgradients) of f_i at the point $\psi_{i-1,k}$. The updates described by Eq. (1.5) are referred to as the *subiterations* of the kth cycle.

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], Gaivoronski [Gai94], Grippo [Gri94], Luo and Tseng [LuT94], Mangasarian and Solodov [MaS94], Bertsekas and Tsitsiklis [BeT96], Bertsekas [Ber97], Tseng [Tse98], Bertsekas and Tsitsiklis [BeT00]). 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]. In the special case where all the stationary points of f are also stationary points of all the component functions f_i, the limit cycle typically reduces to a single point and convergence is obtained; this is the subject of the paper by Solodov [Sol98]. In general, however, the limit cycle consists of m points, each corresponding to one of the subiterations of Eq. (1.5), and these m points are usually distinct.

Incremental subgradient methods exhibit similar behavior to incremental gradient methods, and are similarly motivated by rate of convergence considerations. They were studied first by Kibardin [Kib79], and more recently by Solodov and Zavriev [SoZ98], Nedić and Bertsekas [NeB99], [NeB00], and Ben-Tal, Margalit, and Nemirovski [BMN00]. An asynchronous parallel version of the incremental subgradient method was proposed by Nedić, Bertsekas, and Borkar [NBB00]. Incremental subgradient methods that are somewhat different from the ones of this paper have been proposed by Kaskavelis and Caramanis [KaC98], and Zhao, Luh, and Wang

[ZLW99], while a parallel implementation of related methods was proposed by Kiwiel and Lindberg [KiL00]. These methods share with ours the characteristic of computing a subgradient of only one component f_i per iteration, but differ from ours in that the direction used in an iteration is the sum of the (approximate) subgradients of all the components f_i.

In this paper, we study the convergence properties of the incremental subgradient method for three types of stepsize rules: a *constant stepsize rule*, a *diminishing stepsize rule* (where $\alpha_k \to 0$), and a *dynamic stepsize rule* (where α_k is based on exact or approximate knowledge of the optimal cost function value). Earlier convergence analyses of incremental subgradient methods have focused only on the diminishing stepsize rule. Some understanding into the convergence process is gained by viewing the incremental subgradient method as an approximate subgradient method (or a subgradient method with errors). In particular, we have for all $z \in \Re^n$

$$\begin{aligned}
\left(\sum_{i=1}^m g_{i,k}\right)'(z - x_k) &= \sum_{i=1}^m g'_{i,k}(z - \psi_{i-1,k}) + \sum_{i=1}^m g'_{i,k}(\psi_{i-1,k} - x_k) \\
&\le \sum_{i=1}^m \big(f_i(z) - f_i(\psi_{i-1,k})\big) + \sum_{i=1}^m ||g_{i,k}|| \cdot ||\psi_{i-1,k} - x_k|| \\
&= f(z) - f(x_k) + \sum_{i=2}^m \big(f_i(x_k) - f_i(\psi_{i-1,k})\big) + \sum_{i=2}^m ||g_{i,k}|| \cdot ||\psi_{i-1,k} - x_k|| \\
&\le f(z) - f(x_k) + \sum_{i=2}^m \big(||\tilde g_{i,k}|| + ||g_{i,k}||\big)||\psi_{i-1,k} - x_k|| \\
&\le f(z) - f(x_k) + \sum_{i=2}^m \big(||\tilde g_{i,k}|| + ||g_{i,k}||\big) \left(\alpha_k \sum_{j=1}^{i-1} \|\tilde g_{i,k}\|\right) \\
&\le f(z) - f(x_k) + \epsilon_k,
\end{aligned}$$

where $\tilde g_{i,k} \in \partial f_i(x_k)$, $g_{i,k} \in \partial f_i(\psi_{i-1,k})$, and

$$\epsilon_k = 2\alpha_k \sum_{i=2}^m C_i \left(\sum_{j=1}^{i-1} C_j\right), \qquad C_i = \sup_{k\ge 0}\big\{||g|| \mid g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k})\big\}.$$

Thus if the subgradients $\tilde g_{i,k}$, $g_{i,k}$ are bounded so that the C_i are finite, ϵ_k is bounded and diminishes to zero if $\alpha_k \to 0$. It follows that if a diminishing stepsize rule ($\alpha_k \to 0$) is used and some additional conditions hold, such as $\sum_{k=0}^\infty \alpha_k = \infty$, some of the convergence properties of the incremental method can be derived from known results on ϵ-subgradient methods (see e.g., Dem'yanov and Vasil'ev [DeV85], Polyak [Pol87], p. 144, Correa and Lemaréchal [CoL93], Hiriart-Urruty and Lemaréchal [HiL93], Bertsekas [Ber99]). However, the connection with ϵ-subgradient methods is not helpful for the convergence analysis under the other stepsize rules that we consider (constant and dynamic), because for these rules α_k need not tend to 0, and the same is true for ϵ_k. As a consequence, there are no convergence results for ϵ-subgradient methods under these rules, which can be applied to our analysis.

We also propose a randomized version of the incremental subgradient method (1.4)–(1.6), where the component function f_i in Eq. (1.5) is chosen randomly among the components $f_1, \ldots,$ f_m, according to a uniform distribution. This method may be viewed as a stochastic subgradient method for the problem

$$\min_{x \in X} E_\omega \big\{ f_\omega(x) \big\},$$

where ω is a random variable that is uniformly distributed over the index set $\{1, \ldots, m\}$. Thus some of the insights and analysis from the stochastic subgradient methods can be brought to bear (see e.g., Ermoliev [Erm69], [Erm76], [Erm83], [Erm88], Shor [Sho85] p. 46, Bertsekas and Tsitsiklis [BeT96]). Nonetheless, the idea of using randomization in the context of deterministic nondifferentiable optimization is original and much of our analysis, particularly the part that relates to the constant and the dynamic stepsize rules in Section 3, is also original. An important conclusion, based on Props. 2.1 and 3.1, is that randomization has a significant favorable effect on the method's performance; see also the discussion of Section 3, and Nedić and Bertsekas [NeB99], [NeB00] which provide convergence rate estimates.

The paper is organized as follows. In the next section, we analyze the convergence of the incremental subgradient method under the three types of stepsize rules mentioned above. In Section 3, we establish the convergence properties of randomized versions of the method. Finally, in Section 4, we present some computational results. In particular, we compare the performance of the ordinary subgradient method with that of the incremental subgradient method, and we compare different order rules for processing the component functions f_i within a cycle. The computational results indicate a substantial performance advantage for the randomized processing order over the fixed order. We trace the reason for this to a substantially better error estimate for the randomized order (compare Props. 2.1 and 3.1).

2. CONVERGENCE ANALYSIS OF THE INCREMENTAL SUBGRADIENT METHOD

Throughout this paper, we use the notation

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

where $\|\cdot\|$ denotes the standard Euclidean norm. Our convergence results in this section use the following assumption.

Assumption 2.1: (Subgradient Boundedness) There exist scalars $C_1, \ldots, C_m$ such that

$$||g|| \leq C_i, \qquad \forall\ g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}), \qquad i = 1, \ldots, m, \quad k = 0, 1, \ldots.$$

We note that Assumption 2.1 is satisfied if each f_i is polyhedral (i.e., f_i is the pointwise maximum of a finite number of affine functions). In particular, Assumption 2.1 holds for the dual problem (1.1), (1.2), where for each i and all x the set of subgradients $\partial f_i(x)$ is the convex hull of a finite number of points. More generally, since each component f_i is real-valued and convex over the entire space $\Re^n$, the subdifferential $\partial f_i(x)$ is nonempty and compact for all x and i. If the set X is compact or the sequences $\{\psi_{i,k}\}$ are bounded, then Assumption 2.1 is satisfied since the set $\cup_{x\in B}\partial f_i(x)$ is bounded for any bounded set B (see e.g., Bertsekas [Ber99], Prop. B.24).

The following lemma gives an estimate that will be used repeatedly in the subsequent convergence analysis.

Lemma 2.1: Let Assumption 2.1 hold and let $\{x_k\}$ be the sequence generated by the incremental subgradient method (1.4)–(1.6). Then for all $y \in X$ and $k \geq 0$, we have

$$||x_{k+1} - y||^2 \leq ||x_k - y||^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 C^2, \tag{2.1}$$

where $C = \sum_{i=1}^m C_i$ and C_i is as in Assumption 2.1.

Proof: Using the nonexpansion property of the projection, the subgradient boundedness (cf. Assumption 2.1), and the subgradient inequality for each component function f_i, we obtain for all $y \in X$

$$\begin{aligned}
||\psi_{i,k} - y||^2 &= ||\mathcal{P}_X[\psi_{i-1,k} - \alpha_k g_{i,k}] - y||^2 \\
&\leq ||\psi_{i-1,k} - \alpha_k g_{i,k} - y||^2 \\
&\leq ||\psi_{i-1,k} - y||^2 - 2\alpha_k g_{i,k}'(\psi_{i-1,k} - y) + \alpha_k^2 C_i^2 \\
&\leq ||\psi_{i-1,k} - y||^2 - 2\alpha_k\big(f_i(\psi_{i-1,k}) - f_i(y)\big) + \alpha_k^2 C_i^2, \qquad \forall\ i,k.
\end{aligned}$$

By adding the above inequalities over $i = 1, \ldots, m$, we have for all $y \in X$ and k

$$\begin{aligned}
||x_{k+1} - y||^2 &\leq ||x_k - y||^2 - 2\alpha_k \sum_{i=1}^m \big(f_i(\psi_{i-1,k}) - f_i(y)\big) + \alpha_k^2 \sum_{i=1}^m C_i^2 \\
&= ||x_k - y||^2 - 2\alpha_k \left(f(x_k) - f(y) + \sum_{i=1}^m \big(f_i(\psi_{i-1,k}) - f_i(x_k)\big)\right) + \alpha_k^2 \sum_{i=1}^m C_i^2.
\end{aligned}$$

By strengthening the above inequality, we have for all $y \in X$ and k

$$\begin{aligned}
||x_{k+1} - y||^2 &\leq ||x_k - y||^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + 2\alpha_k \sum_{i=1}^m C_i||\psi_{i-1,k} - x_k|| + \alpha_k^2 \sum_{i=1}^m C_i^2 \\
&\leq ||x_k - y||^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 \left(2\sum_{i=2}^m C_i \left(\sum_{j=1}^{i-1} C_j\right) + \sum_{i=1}^m C_i^2\right) \\
&= ||x_k - y||^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 \left(\sum_{i=1}^m C_i\right)^2 \\
&= ||x_k - y||^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 C^2,
\end{aligned}$$

where in the first inequality we use the relation

$$f_i(x_k) - f_i(\psi_{i-1,k}) \leq ||\tilde{g}_{i,k}|| \cdot ||\psi_{i-1,k} - x_k|| \leq C_i||\psi_{i-1,k} - x_k||$$

with $\tilde{g}_{i,k} \in \partial f_i(x_k)$, and in the second inequality we use the relation

$$||\psi_{i,k} - x_k|| \leq \alpha_k \sum_{j=1}^{i} C_j, \qquad i = 1, \ldots, m, \quad k \geq 0,$$

which follows from Eqs. (1.4)–(1.6) and Assumption 2.1. **Q.E.D.**

Among other things, Lemma 2.1 guarantees that given the current iterate x_k and some other point $y \in X$ with lower cost than x_k, the next iterate x_{k+1} will be closer to y than x_k, provided the stepsize α_k is sufficiently small [less than $2\big(f(x_k) - f(y)\big)/C^2$]. This fact is used repeatedly, with a variety of choices for y, in what follows.

Constant Stepsize Rule

We first consider the case of a constant stepsize rule.

Proposition 2.1: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental method (1.4)–(1.6) with the stepsize α_k fixed to some positive constant α, we have:

(a) If $f^* = -\infty$, then

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

(b) If $f^* > -\infty$, then

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

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

Proof: We prove (a) and (b) simultaneously. If the result does not hold, there must exist an $\epsilon > 0$ such that

$$\liminf_{k\to\infty} f(x_k) > f^* + \frac{\alpha C^2}{2} + 2\epsilon.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k\to\infty} f(x_k) \geq f(\hat{y}) + \frac{\alpha C^2}{2} + 2\epsilon,$$

and let k_0 be large enough so that for all $k \geq k_0$ we have

$$f(x_k) \geq \liminf_{k\to\infty} f(x_k) - \epsilon.$$

By adding the preceding two relations, we obtain for all $k \geq k_0$

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha C^2}{2} + \epsilon.$$

Using Lemma 2.1 for the case where $y = \hat{y}$ together with the above relation, we obtain for all $k \geq k_0$, $||x_{k+1} - \hat{y}||^2 \leq ||x_k - \hat{y}||^2 - 2\alpha\epsilon$. Thus we have

$$||x_{k+1} - \hat{y}||^2 \leq ||x_k - \hat{y}||^2 - 2\alpha\epsilon \leq ||x_{k-1} - \hat{y}||^2 - 4\alpha\epsilon \leq \cdots \leq ||x_{k_0} - \hat{y}||^2 - 2(k + 1 - k_0)\alpha\epsilon,$$

which cannot hold for k sufficiently large – a contradiction. **Q.E.D.**

Diminishing Stepsize Rule

The next result is the analog of a classical convergence result for the ordinary subgradient method of Ermoliev [Erm66] (see also Polyak [Pol67]).

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 method (1.4)–(1.6), we have

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

Proof: Use Lemma 2.1 with $y \in X^*$ and Prop. 1.2 of Correa and Lemaréchal [CoL93]. **Q.E.D.**

If we assume in addition that X^* is nonempty and bounded, Prop. 2.2 can be strengthened as in the next proposition. This proposition is similar to a result of Solodov and Zavriev [SoZ98], which was proved by different methods under the stronger assumption that X is a compact set.

Proposition 2.3: Let Assumption 2.1 hold, and let X^* be nonempty and bounded. Also, 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 (1.4)–(1.6), we have

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

Proof: The idea is to show that once x_k enters a certain level set, it cannot get too far away from that set. Fix a $\gamma > 0$, and let k_0 be such that $\gamma \geq \alpha_k C^2$ for all $k \geq k_0$. We distinguish two cases:

Case 1: $f(x_k) > f^* + \gamma$. From Lemma 2.1 we obtain for all $x^* \in X^*$ and k

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - 2\alpha_k\big(f(x_k) - f^*\big) + \alpha_k^2 C^2. \tag{2.2}$$

Hence

$$||x_{k+1} - x^*||^2 < ||x_k - x^*||^2 - 2\gamma\alpha_k + \alpha_k^2 C^2 = ||x_k - x^*||^2 - \alpha_k(2\gamma - \alpha_k C^2) \leq ||x_k - x^*||^2 - \alpha_k\gamma,$$

so that

$$dist(x_{k+1}, X^*) \leq dist(x_k, X^*) - \alpha_k\gamma. \tag{2.3}$$

Case 2: $f(x_k) \leq f^* + \gamma$. This case must occur for infinitely many k, in view of Eq. (2.3) and the fact $\sum_{k=0}^{\infty} \alpha_k = \infty$. Since x_k belongs to the level set

$$L_\gamma = \big\{y \in X \mid f(y) \leq f^* + \gamma\big\},$$

which is bounded (in view of the boundedness of X^*), we have

$$dist(x_k, X^*) \leq d(\gamma) < \infty, \tag{2.4}$$

where we denote

$$d(\gamma) = \max_{y \in L_\gamma} dist(y, X^*).$$

From the iteration (1.4)–(1.6), we have $||x_{k+1} - x_k|| \leq \alpha_k C$, so for all $x^* \in X^*$

$$||x_{k+1} - x^*|| \leq ||x_k - x^*|| + ||x_{k+1} - x_k|| \leq ||x_k - x^*|| + \alpha_k C.$$

By taking the minimum above over $x^* \in X^*$ and by using Eq. (2.4), we obtain

$$dist(x_{k+1}, X^*) \leq d(\gamma) + \alpha_k C. \tag{2.5}$$

Combining Eq. (2.3) which holds when $f(x_k) > f^* + \gamma$ (Case 1 above), with Eq. (2.5) which holds for the infinitely many k for which $f(x_k) \leq f^* + \gamma$ (Case 2 above), we see that

$$dist(x_k, X^*) \leq d(\gamma) + \alpha_k C, \qquad \forall\ k \geq k_0.$$

Therefore, since $\alpha_k \to 0$,

$$\limsup_{k \to \infty} dist(x_k, X^*) \leq d(\gamma), \qquad \forall\ \gamma > 0.$$

Since, in view of the continuity of f and the compactness of its level sets, we have $\lim_{\gamma \to 0} d(\gamma) = 0$, we obtain $\lim_{k \to \infty} dist(x_k, X^*) = 0$. This relation also implies that $\lim_{k \to \infty} f(x_k) = f^*$. **Q.E.D.**

The assumption that X^* is nonempty and bounded holds, for example, if all $\inf_{x \in X} f_i(x)$ are finite and at least one of the components f_i has bounded level sets (see Rockafellar [Roc70], Theorem 9.3). Proposition 2.3 does not guarantee convergence of the entire sequence $\{x_k\}$. With

slightly different assumptions that include an additional mild restriction on the stepsize sequence, this convergence is guaranteed, as indicated in the following proposition.

Proposition 2.4: Let Assumption 2.1 hold and let the optimal set X^* be nonempty. Also 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 $\{x_k\}$ generated by the incremental subgradient method (1.4)–(1.6) converges to some optimal solution.

Proof: Use Lemma 2.1 with $y \in X^*$ and Prop. 1.3 of Correa and Lemaréchal [CoL93]. **Q.E.D.**

In Props. 2.2–2.4, we use the same stepsize α_k in all subiterations of a cycle. As shown by Kibardin in [Kib79] and by Nedić, Bertsekas, and Borkar in [NBB00] (for a more general incremental method), the convergence can be preserved if we vary the stepsize α_k within each cycle, provided that the variations of α_k in the cycles are suitably small.

Dynamic Stepsize Rule for Known 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

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{||g_k||^2},$$

with $g_k \in \partial f(x_k)$, $0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2$, introduced by Polyak in [Pol69] (see also discussions in Shor [Sho85], Brännlund [Brä93], and Bertsekas [Ber99]). For the incremental method, we propose a variant of this stepsize where $||g_k||$ is replaced by an upper bound C:

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{C^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \tag{2.6}$$

where

$$C = \sum_{i=1}^{m} C_i \tag{2.7}$$

and

$$C_i \geq \sup_{k \geq 0} \big\{ ||g|| \mid g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}) \big\}, \qquad i = 1, \ldots, m. \tag{2.8}$$

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

We first consider the case where f^* is known. We later modify the stepsize, so that f^* can be replaced by a dynamically updated estimate.

Proposition 2.5: Let Assumption 2.1 hold and let the optimal set X^* be nonempty. Then the sequence $\{x_k\}$ generated by the incremental subgradient method (1.4)–(1.6) with the dynamic stepsize rule (2.6)–(2.8) converges to some optimal solution.

Proof: From Lemma 2.1 with $y = x^* \in X^*$, we have

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - 2\alpha_k\big(f(x_k) - f^*\big) + \alpha_k^2 C^2, \qquad \forall\, x^* \in X^*, \quad k \geq 0,$$

and by using the definition of α_k [cf. Eq. (2.6)], we obtain

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - \underline{\gamma}(2 - \overline{\gamma})\frac{\big(f(x_k) - f^*\big)^2}{C^2}, \qquad \forall\, x^* \in X^*, \quad k \geq 0.$$

This implies that $\{x_k\}$ is bounded. Furthermore, $f(x_k) \to f^*$, since otherwise we would have $||x_{k+1} - x^*|| \leq ||x_k - x^*|| - \epsilon$ for some suitably small $\epsilon > 0$ and infinitely many k. Hence for any limit point $\overline{x}$ of $\{x_k\}$, we have $\overline{x} \in X^*$, and since the sequence $\{||x_k - x^*||\}$ is decreasing, it converges to $||\overline{x} - x^*||$ for every $x^* \in X^*$. If there are two distinct limit points $\tilde{x}$ and $\overline{x}$ of $\{x_k\}$, we must have $\tilde{x} \in X^*$, $\overline{x} \in X^*$, and $||\tilde{x} - x^*|| = ||\overline{x} - x^*||$ for all $x^* \in X^*$, which is possible only if $\tilde{x} = \overline{x}$. **Q.E.D.**

Dynamic Stepsize Rule for Unknown f^*

In most practical problems the value f^* is not known. In this case we may modify the dynamic stepsize (2.6) by replacing f^* with an estimate. This leads to the stepsize rule

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{C^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \qquad \forall\, k \geq 0, \tag{2.9}$$

where C is defined by Eqs. (2.7), (2.8), and $f_k^{\rm lev}$ is an estimate of f^*.

We discuss two procedures for updating $f_k^{\rm lev}$. In both procedures $f_k^{\rm lev}$ is equal to the best function value $\min_{0\leq j\leq k} f(x_j)$ achieved up to the kth iteration minus a positive amount δ_k which is adjusted based on the algorithm's progress. The first adjustment procedure (new even when specialized to the ordinary subgradient method) 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 Brännlund [Brä93], and Goffin and Kiwiel [GoK99].

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

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta_k, \tag{2.10}$$

and δ_k is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k & \text{if } f(x_{k+1}) \leq f_k^{\text{lev}}, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\text{lev}}, \end{cases} \tag{2.11}$$

where δ, β, and ρ are fixed positive constants with $\beta < 1$ and $\rho \geq 1$. 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. (2.9) is bounded away from zero, since from Eq. (2.10), we have $f(x_k) - f_k^{\text{lev}} \geq \delta$ and hence

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

As a result, the method behaves similar to the one with a constant stepsize (cf. Prop. 2.1), as indicated by the following proposition.

Proposition 2.6: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental method (1.4)–(1.6) and the dynamic stepsize rule (2.9) with the adjustment procedure (2.10)–(2.11), we have:

(a) If $f^* = -\infty$, then

$$\inf_{k\geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

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

Proof: To arrive at a contradiction, assume that

$$\inf_{k\geq 0} f(x_k) > f^* + \delta. \tag{2.12}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\text{lev}}$], the current best function value $\min_{0\leq j\leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (2.10) and (2.11)], so in view of Eq. (2.12), the target value can be attained only a finite number times. From Eq. (2.11) it follows that after finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is an index $\overline{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \overline{k}. \tag{2.13}$$

In view of Eq. (2.12), there exists $\overline{y} \in X$ such that $\inf_{k\geq 0} f(x_k) - \delta \geq f(\overline{y})$. From Eqs. (2.10) and (2.13), we have

$$f_k^{\text{lev}} = \min_{0\leq j\leq k} f(x_j) - \delta \geq \inf_{k\geq 0} f(x_k) - \delta \geq f(\overline{y}), \qquad \forall\ k \geq \overline{k},$$

so that

$$\alpha_k\big(f(x_k)-f(\overline{y})\big)\geq\alpha_k\big(f(x_k)-f_k^{\rm lev}\big)=\gamma_k\left(\frac{f(x_k)-f_k^{\rm lev}}{C}\right)^2,\qquad\forall\ k\geq\overline{k}.$$

By using Lemma 2.1 with $y=\overline{y}$, we have

$$||x_{k+1}-\overline{y}||^2\leq||x_k-\overline{y}||^2-2\alpha_k\big(f(x_k)-f(\overline{y})\big)+\alpha_k^2C^2,\qquad\forall\ k\geq0.$$

By combining the preceding two relations and the definition of α_k [cf. Eq. (2.9)], we obtain

$$\begin{aligned}||x_{k+1}-\overline{y}||^2&\leq||x_k-\overline{y}||^2-2\gamma_k\left(\frac{f(x_k)-f_k^{\rm lev}}{C}\right)^2+\gamma_k^2\left(\frac{f(x_k)-f_k^{\rm lev}}{C}\right)^2\\&=||x_k-\overline{y}||^2-\gamma_k(2-\gamma_k)\left(\frac{f(x_k)-f_k^{\rm lev}}{C}\right)^2\\&\leq||x_k-\overline{y}||^2-\underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{C^2},\qquad\forall\ k\geq\overline{k},\end{aligned}$$

where the last inequality follows from the facts $\gamma_k\in[\underline{\gamma},\overline{\gamma}]$ and $f(x_k)-f_k^{\rm lev}\geq\delta$ for all k. By summing the above inequalities over k, we have

$$||x_k-\overline{y}||^2\leq||x_{\overline{k}}-\overline{y}||^2-(k-\overline{k})\underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{C^2},\qquad\forall\ k\geq\overline{k},$$

which cannot hold for large k – a contradiction. **Q.E.D.**

When $m=1$, the incremental subgradient method (1.4)–(1.6) becomes the ordinary subgradient method

$$x_{k+1}=\mathcal{P}_X[x_k-\alpha_kg_k],\qquad\forall\ k\geq0.$$

The dynamic stepsize rule (2.9) using the adjustment procedure of Eqs. (2.10)–(2.11) (with $C=||g_k||$), and the convergence result of Prop. 2.6 are new to our knowledge for this method.

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.

Path-Based 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., $f_k^{\rm rec}=\min_{0\leq j\leq k}f(x_j)$].

Step 2 (*Sufficient descent*) If $f(x_k) \leq f^{\rm rec}_{k(l)} - \frac{\delta_l}{2}$, 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 $f^{\rm lev}_k = f^{\rm rec}_{k(l)} - \delta_l$. Select $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and calculate x_{k+1} via Eqs. (1.4)–(1.6) with the stepsize (2.9).

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

The algorithm uses the same target level $f^{\rm lev}_k = f^{\rm rec}_{k(l)} - \delta_l$ for $k = k(l), k(l)+1, \ldots, k(l+1)-1$. The target level is updated only if sufficient descent or oscillation is detected (Step 2 or Step 3, respectively). It can be shown that the value σ_k is an upper bound on the length of the path traveled by iterates $x_{k(l)}, \ldots, x_k$ for $k < k(l+1)$. Whenever σ_k exceeds the prescribed upper bound B on the path length, the parameter δ_l is decreased, which (possibly) increases the target level $f^{\rm lev}_k$.

We will show that $\inf_{k\geq 0} f(x_k) = f^*$ even if f^* is not finite. First, we give a preliminary result showing that the target values $f^{\rm lev}_k$ are updated infinitely often (i.e. $l \to \infty$), and that $\inf_{k\geq 0} f(x_k) = -\infty$ if δ_l is nondiminishing.

Lemma 2.2: Let Assumption 2.1 hold. Then for the path-based incremental target level algorithm we have $l \to \infty$, and either $\inf_{k\geq 0} f(x_k) = -\infty$ or $\lim_{l\to\infty} \delta_l = 0$.

Proof: Assume that l takes only a finite number of values, say $l = 0, 1, \ldots, \overline{l}$. In this case we have $\sigma_k + \alpha_k C = \sigma_{k+1} \leq B$ for all $k \geq k(\overline{l})$, so that $\lim_{k\to\infty} \alpha_k = 0$. But this is impossible, since for all $k \geq k(\overline{l})$ we have

$$\alpha_k = \gamma_k \frac{f(x_k) - f^{\rm lev}_k}{C^2} \geq \underline{\gamma}\, \frac{\delta_{\overline{l}}}{C^2} > 0.$$

Hence $l \to \infty$.

Let $\delta = \lim_{l\to\infty} \delta_l$. If $\delta > 0$, then from Steps 2 and 3 it follows that for all l large enough, we h ave $\delta_l = \delta$ and

$$f^{\rm rec}_{k(l+1)} - f^{\rm rec}_{k(l)} \leq -\frac{\delta}{2},$$

implying that $\inf_{k\geq 0} f(x_k) = -\infty$. **Q.E.D.**

We have the following convergence result. In the special case of the ordinary subgradient method, this result was proved by Goffin and Kiwiel [GoK99] using a different (and much longer) proof.

Proposition 2.7: Let Assumption 2.1 hold. Then, for the path-based incremental target level algorithm, we have

$$\inf_{k\geq 0} f(x_k) = f^*.$$

Proof: If $\lim_{l\to\infty} \delta_l > 0$, then, according to Lemma 2.2, we have $\inf_{k\geq 0} f(x_k) = -\infty$ and we are done, so assume that $\lim_{l\to\infty} \delta_l = 0$. Let L be given by

$$L = \left\{ l \in \{1, 2, \ldots\} \;\middle|\; \delta_l = \frac{\delta_{l-1}}{2} \right\}.$$

Then, from Steps 3 and 5, we obtain

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} C = \sum_{j=k(l)}^{k-1} C\alpha_j,$$

so that $k(l+1) = k$ and $l+1 \in L$ whenever $\sum_{j=k(l)}^{k-1} \alpha_j C > B$ at Step 3. Hence

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \frac{B}{C}, \qquad \forall\, l \in L,$$

and, since the cardinality of L is infinite, we have

$$\sum_{k=k(\hat{l})}^{\infty} \alpha_k \geq \sum_{l\geq \hat{l},\ l\in L} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \sum_{l\geq \hat{l},\ l\in L} \frac{B}{C} = \infty. \tag{2.14}$$

Now, in order to arrive at a contradiction, assume that $\inf_{k\geq 0} f(x_k) > f^*$, so that for some $\hat{y} \in X$ and some $\epsilon > 0$

$$\inf_{k\geq 0} f(x_k) - \epsilon \geq f(\hat{y}). \tag{2.15}$$

Since $\delta_l \to 0$, there is a large enough $\hat{l}$ such that $\delta_l \leq \epsilon$ for all $l \geq \hat{l}$, so that for all $k \geq k(\hat{l})$

$$f_k^{\text{lev}} = f_{k(l)}^{\text{rec}} - \delta_l \geq \inf_{k\geq 0} f(x_k) - \epsilon \geq f(\hat{y}).$$

Using this relation, Lemma 2.1 for $y = \hat{y}$, and the definition of α_k, we obtain

$$\begin{aligned}
||x_{k+1} - \hat{y}||^2 &\leq ||x_k - \hat{y}||^2 - 2\alpha_k \big(f(x_k) - f(\hat{y})\big) + \alpha_k^2 C^2. \\
&\leq ||x_k - \hat{y}||^2 - 2\alpha_k \big(f(x_k) - f_k^{\text{lev}}\big) + \alpha_k^2 C^2 \\
&= ||x_k - \hat{y}||^2 - \gamma_k(2-\gamma_k) \frac{\big(f(x_k) - f_k^{\text{lev}}\big)^2}{C^2} \\
&\leq ||x_k - \hat{y}||^2 - \underline{\gamma}(2-\overline{\gamma}) \frac{\big(f(x_k) - f_k^{\text{lev}}\big)^2}{C^2}.
\end{aligned}$$

By summing these inequalities over $k \geq k(\hat{l})$, we have

$$\frac{\underline{\gamma}(2-\overline{\gamma})}{C^2} \sum_{k=k(\hat{l})}^{\infty} \big(f(x_k) - f_k^{\text{lev}}\big)^2 \leq ||x_{k(\hat{l})} - \hat{y}||^2,$$

and consequently $\sum_{k=k(\hat{l})}^{\infty} \alpha_k^2 < \infty$ [see the definition of α_k in Eq. (2.9)]. Since $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$ [cf. Eq. (2.14)], according to Prop. 2.2, we must have

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

Hence $\inf_{k\geq 0} f(x_k) = f^*$, which contradicts Eq. (2.15). **Q.E.D.**

In an attempt to improve the efficiency of the path-based incremental target level algorithm, one may introduce parameters $\beta, \tau \in (0,1)$ and $\rho \geq 1$ (whose values will be fixed at Step 0), and modify Steps 2 and 3 as follows:

Step 2′ If $f(x_k) \leq f_{k(l)}^{\rm rec} - \tau\delta_l$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \rho\delta_l$, increase l by 1, and go to Step 4.

Step 3′ If $\sigma_k > B$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \beta\delta_l$, and increase l by 1.

It can be seen that the result of Prop. 2.7 still holds for this modified algorithm. If we choose $\rho > 1$ at Step 3′, then in the proofs of Lemma 2.2 and Prop. 2.7 we have to replace $\lim_{l\to\infty} \delta_l$ with $\limsup_{l\to\infty} \delta_l$.

Let us remark that there is no need to keep the path bound B fixed. Instead, we can decrease B as the method progresses. However, in order to preserve the convergence result of Prop. 2.7, we have to ensure that B remains bounded away from zero for otherwise $f_k^{\rm rec}$ might converge to a nonoptimal value.

It can be verified that all the results presented in this section are valid for the incremental method that does not use projections within the cycles, but rather employs projections at the end of cycles:

$$\psi_{i,k} = \psi_{i-1,k} - \alpha_k g_{i,k}, \qquad g_{i,k} \in \partial f_i(\psi_{i-1,k}), \quad i = 1, \ldots, m,$$

where $\psi_{0,k} = x_k$ and the iterate x_{k+1} is given by

$$x_{k+1} = \mathcal{P}_X[\psi_{m,k}].$$

This method and its modifications, including additive-type errors on subgradients, synchronous parallelization, and a momentum term is given by Solodov and Zavriev [SoZ98], and is analyzed for the case of a compact set X and a diminishing stepsize rule.

3. AN INCREMENTAL SUBGRADIENT METHOD WITH RANDOMIZATION

It can be verified that the preceding convergence analysis goes through assuming any order for processing the component functions f_i, as long as each component is taken into account

exactly once within a cycle. In particular, at the beginning of each cycle k, we could reorder the components f_i by either shifting or reshuffling and then proceed with the calculations until the end of the cycle. However, the order used can significantly affect the rate of convergence of the method. Unfortunately, determining the most favorable order 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 (see also [BeT00]). Similarly, the corresponding incremental subgradient method at each step picks randomly a function f_i to be processed next. For the case of a diminishing stepsize, the convergence of the method follows from known stochastic subgradient convergence results (e.g., Ermoliev [Erm69], [Erm88], Polyak [Pol87], p. 159) – see the subsequent Prop. 3.2. In this section, we also analyze the method for the constant and dynamic stepsize rules. This analysis is new and has no counterpart in the available stochastic subgradient literature.

The formal description of the randomized method is as follows

$$x_{k+1} = \mathcal{P}_X\big[x_k - \alpha_k g(\omega_k, x_k)\big], \tag{3.1}$$

where ω_k is a random variable taking equiprobable values from the set $\{1, \ldots, m\}$, and $g(\omega_k, x_k)$ is a subgradient of the component f_{ω_k} at x_k. This simply means that if the random variable ω_k takes a value j, then the vector $g(\omega_k, x_k)$ is a subgradient of f_j at x_k.

Throughout this section we assume the following regarding the randomized method (3.1).

Assumption 3.1:

(a) The sequence $\{\omega_k\}$ is a sequence of independent random variables, each uniformly distributed over the set $\{1, \ldots, m\}$. Furthermore, the sequence $\{\omega_k\}$ is independent of the sequence $\{x_k\}$.

(b) The set of subgradients $\big\{g(\omega_k, x_k) \mid k = 0, 1, \ldots\big\}$ is bounded, i.e., there exists a positive constant C_0 such that with probability 1

$$||g(\omega_k, x_k)|| \leq C_0, \qquad \forall\ k \geq 0.$$

Note that if the set X is compact or the components f_i are polyhedral, then Assumption 3.1(b) is satisfied. The proofs of several propositions in this section rely on the supermartingale convergence theorem as stated for example in Bertsekas and Tsitsiklis [BeT96], p. 148.

Theorem 3.1: (Supermartingale Convergence Theorem) Let Y_k, Z_k, and W_k, $k = 0, 1, 2, \ldots$, be three sequences of random variables and let $\mathcal{F}_k$, $k = 0, 1, 2, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:

(a) The random variables Y_k, Z_k, and W_k are nonnegative, and are functions of the random variables in $\mathcal{F}_k$.

(b) For each k, we have $E\{Y_{k+1} \mid \mathcal{F}_k\} \leq Y_k - Z_k + W_k$.

(c) There holds $\sum_{k=0}^{\infty} W_k < \infty$.

Then, we have $\sum_{k=0}^{\infty} Z_k < \infty$, and the sequence Y_k converges to a nonnegative random variable Y, with probability 1.

Constant Stepsize Rule

Proposition 3.1: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the randomized incremental method (3.1), with the stepsize α_k fixed to some positive constant α, we have:

(a) If $f^* = -\infty$, then with probability 1

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then with probability 1

$$\inf_{k \geq 0} f(x_k) \leq f^* + \frac{\alpha m C_0^2}{2}.$$

Proof: By adapting Lemma 2.1 to the case where f is replaced by f_{ω_k}, we have

$$||x_{k+1} - y||^2 \leq ||x_k - y||^2 - 2\alpha\big(f_{\omega_k}(x_k) - f_{\omega_k}(y)\big) + \alpha^2 C_0^2, \qquad \forall\ y \in X, \quad k \geq 0.$$

By taking the conditional expectation with respect to $\mathcal{F}_k = \{x_0, \ldots, x_k\}$, the method's history up to x_k, we obtain for all $y \in X$ and k

$$\begin{aligned} E\big\{||x_{k+1} - y||^2 \mid \mathcal{F}_k\big\} &\leq ||x_k - y||^2 - 2\alpha E\big\{f_{\omega_k}(x_k) - f_{\omega_k}(y) \mid \mathcal{F}_k\big\} + \alpha^2 C_0^2 \\ &= ||x_k - y||^2 - 2\alpha \sum_{i=1}^{m} \frac{1}{m}\big(f_i(x_k) - f_i(y)\big) + \alpha^2 C_0^2 \\ &= ||x_k - y||^2 - \frac{2\alpha}{m}\big(f(x_k) - f(y)\big) + \alpha^2 C_0^2, \end{aligned} \tag{3.2}$$

where the first equality follows since ω_k takes the values $1, \ldots, m$ with equal probability $1/m$.

Now, fix a nonnegative integer N, consider the level set L_N defined by

$$L_N = \begin{cases} \left\{x \in X \mid f(x) < -N + 1 + \frac{\alpha m C_0^2}{2}\right\} & \text{if } f^* = -\infty, \\ \left\{x \in X \mid f(x) < f^* + \frac{2}{N} + \frac{\alpha m C_0^2}{2}\right\} & \text{if } f^* > -\infty, \end{cases}$$

and let $y_N \in X$ be such that

$$f(y_N) = \begin{cases} -N & \text{if } f^* = -\infty, \\ f^* + \frac{1}{N} & \text{if } f^* > -\infty. \end{cases}$$

Note that $y_N \in L_N$ by construction. Define a new process $\{\hat{x}_k\}$ as follows

$$\hat{x}_{k+1} = \begin{cases} \mathcal{P}_X\big[\hat{x}_k - \alpha g(\omega_k, \hat{x}_k)\big] & \text{if } \hat{x}_k \notin L_N, \\ y_N & \text{otherwise,} \end{cases}$$

where $\hat{x}_0 = x_0$. Thus the process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once x_k enters the level set L_N, the process terminates with $\hat{x}_k = y_N$ (since $y_N \in L_N$). Using Eq. (3.2) with $y = y_N$, we have

$$E\big\{||\hat{x}_{k+1} - y_N||^2 \mid \mathcal{F}_k\big\} \le ||\hat{x}_k - y_N||^2 - \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) + \alpha^2 C_0^2,$$

or equivalently

$$E\big\{||\hat{x}_{k+1} - y_N||^2 \mid \mathcal{F}_k\big\} \le ||\hat{x}_k - y_N||^2 - z_k, \tag{3.3}$$

where

$$z_k = \begin{cases} \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) - \alpha^2 C_0^2 & \text{if } \hat{x}_k \notin L_N, \\ 0 & \text{if } \hat{x}_k = y_N. \end{cases}$$

(a) Let $f^* = -\infty$. Then if $\hat{x}_k \notin L_N$, we have

$$z_k = \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) - \alpha^2 C_0^2 \ge \frac{2\alpha}{m}\left(-N + 1 + \frac{\alpha m C_0^2}{2} + N\right) - \alpha^2 C_0^2 = \frac{2\alpha}{m}.$$

Since $z_k = 0$ for $\hat{x}_k \in L_N$, we have $z_k \ge 0$ for all k, and by Eq. (3.3) and the supermartingale convergence theorem, $\sum_{k=0}^{\infty} z_k < \infty$ implying that $\hat{x}_k \in L_N$ for sufficiently large k, with probability 1. Therefore, in the original process we have

$$\inf_{k \ge 0} f(x_k) \le -N + 1 + \frac{\alpha m C_0^2}{2}$$

with probability 1. Letting $N \to \infty$, we obtain $\inf_{k \ge 0} f(x_k) = -\infty$ with probability 1.

(b) Let $f^* > -\infty$. Then if $\hat{x}_k \notin L_N$, we have

$$z_k = \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) - \alpha^2 C_0^2 \ge \frac{2\alpha}{m}\left(f^* + \frac{2}{N} + \frac{\alpha m C_0^2}{2} - f^* - \frac{1}{N}\right) - \alpha^2 C_0^2 = \frac{2\alpha}{mN}.$$

Hence, $z_k \geq 0$ for all k, and by the supermartingale convergence theorem, we have $\sum_{k=0}^{\infty} z_k < \infty$ implying that $\hat{x}_k \in L_N$ for sufficiently large k, so that in the original process,

$$\inf_{k\geq 0} f(x_k) \leq f^* + \frac{2}{N} + \frac{\alpha m C_0^2}{2}$$

with probability 1. Letting $N \to \infty$, we obtain $\inf_{k\geq 0} f(x_k) \leq f^* + \alpha m C_0^2/2$. **Q.E.D.**

From Prop. 3.1(b), it can be seen that when $f^* > -\infty$, the randomized method (3.1) with a fixed stepsize has a better error bound (by a factor m, since $C^2 \approx m^2 C_0^2$) than the one of the nonrandomized method (1.4)–(1.6) with the same stepsize (cf. Prop. 2.1). This indicates that when randomization is used, the stepsize α_k should generally be chosen larger than in the nonrandomized methods of Section 2. This can also be observed from our experimental results. Being able to use a larger stepsize suggests a potential rate of convergence advantage in favor of the randomized methods, which is consistent with our experimental results. A more precise result is shown in Nedić and Bertsekas [NeB00]: given any $\epsilon > 0$, by using $m\big(dist(x_0, X^*)\big)^2/\alpha\epsilon$ iterations of the nonrandomized method, we are guaranteed a cost function value that is within a tolerance $(\alpha m^2 C_0^2 + \epsilon)/2$ from the optimum f^*, while by using the same *expected* number of iterations of the randomized method, we are guaranteed a cost function value that is within the potentially much smaller tolerance $(\alpha m C_0^2 + \epsilon)/2$ from f^*.

Diminishing Stepsize Rule

As mentioned earlier, the randomized method (3.1) with a diminishing stepsize can be viewed as a special case of a stochastic subgradient method. Consequently, we just state the main convergence result and refer to the literature for its proof.

Proposition 3.2: Let Assumption 3.1 hold and let the optimal set X^* be nonempty. Also assume that the stepsize α_k in Eq. (3.1) 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 $\{x_k\}$ generated by the randomized method (3.1) converges to some optimal solution with probability 1.

Proof: Similar to the proof of Prop. 3.1, we obtain for all k and $x^* \in X^*$

$$E\big\{||x_{k+1} - x^*||^2 \mid \mathcal{F}_k\big\} \leq ||x_k - x^*||^2 - \frac{2\alpha_k}{m}\big(f(x_k) - f^*\big) + \alpha_k^2 C^2$$

[cf. Eq. (3.2) with α and y replaced by α_k and x^*, respectively], where $\mathcal{F}_k = \{x_0, \ldots, x_k\}$. By the supermartingale convergence theorem, for each $x^* \in X^*$, we have for all sample paths in a

set Ω_{x^*} of probability 1

$$\sum_{k=0}^{\infty} \frac{2\alpha_k}{m}\big(f(x_k) - f^*\big) < \infty, \tag{3.4}$$

and that $\{\|x_k - x^*\|\}$ converges.

Let $\{v_i\}$ be a countable subset of the relative interior $\mathrm{ri}(X^*)$ that is dense in X^* [such a set exists since $\mathrm{ri}(X^*)$ is a relatively open subset of the affine hull of X^*; an example of such a set is the intersection of X^* with the set of vectors of the form $x^* + \sum_{i=1,}^{p} r_i\xi_i$, where $\xi_1, \ldots, \xi_p$ are basis vectors for the affine hull of X^* and r_i are rational numbers]. The intersection

$$\Omega = \cap_{i=1}^{\infty}\Omega_{v_i}$$

has probability 1, since its complement $\overline{\Omega}$ is equal to $\cup_{i=1}^{\infty}\overline{\Omega_{v_i}}$ and

$$\mathrm{Prob}\left(\cup_{i=1}^{\infty}\overline{\Omega_{v_i}}\right) \leq \sum_{i=1}^{\infty}\mathrm{Prob}\left(\overline{\Omega_{v_i}}\right) = 0.$$

For each sample path in Ω, the sequences $\|x_k - v_i\|$ converge so that $\{x_k\}$ is bounded. Furthermore, from Eq. (3.4) we see that

$$\lim_{k\to\infty}\big(f(x_k) - f^*\big) = 0,$$

which by continuity of f implies that all the limit points of $\{x_k\}$ belong to X^*. Since $\{v_i\}$ is a dense subset of X^* and the sequences $\|x_k - v_i\|$ converge, it follows that $\{x_k\}$ cannot have more than one limit point, so it must converge to some vector $\overline{x} \in X^*$. **Q.E.D.**

Dynamic Stepsize Rule for Known f^*

One possible version of the dynamic stepsize rule for the method (3.1) has the form

$$\alpha_k = \gamma_k\frac{f(x_k) - f^*}{mC_0^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2,$$

where $\{\gamma_k\}$ is a deterministic sequence, and requires knowledge of the cost function value $f(x_k)$ at the current iterate x_k. However, it would be inefficient to compute $f(x_k)$ at each iteration since that iteration involves a single component f_i, while the computation of $f(x_k)$ requires all the components. We thus modify the dynamic stepsize rule so that the value of f and the parameter γ_k that are used in the stepsize formula are updated every M iterations, where M is any fixed positive integer, rather than at each iteration. In particular, assuming f^* is known, we use the stepsize

$$\alpha_k = \gamma_p\frac{f(x_{Mp}) - f^*}{mMC_0^2}, \quad 0 < \underline{\gamma} \leq \gamma_p \leq \overline{\gamma} < 2, \quad k = Mp, \ldots, M(p+1) - 1, \quad p = 0, 1, \ldots, \tag{3.5}$$

where $\{\gamma_p\}$ is a deterministic sequence. We can choose M greater than m, if m is relatively small, or we can select M smaller than m, if m is very large.

Proposition 3.3: Let Assumption 3.1 hold and let X^* be nonempty. Then the sequence $\{x_k\}$ generated by the randomized method (3.1) with the stepsize (3.5) converges to some optimal solution with probability 1.

Proof: By adapting Lemma 2.1 to the case where $y = x^* \in X^*$ and f is replaced by f_{ω_k}, we have

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - 2\alpha_k\big(f_{\omega_k}(x_k) - f_{\omega_k}(x^*)\big) + \alpha_k^2 C_0^2, \qquad \forall\, x^* \in X^*, \quad k \geq 0.$$

By summing this inequality over $k = Mp, \ldots, M(p+1)-1$ (i.e., over the M iterations of a cycle), we obtain for all $x^* \in X^*$ and all p

$$||x_{M(p+1)} - x^*||^2 \leq ||x_{Mp} - x^*||^2 - 2\alpha_{Mp} \sum_{k=Mp}^{M(p+1)-1} \big(f_{\omega_k}(x_k) - f_{\omega_k}(x^*)\big) + M\alpha_{Mp}^2 C_0^2,$$

since $\alpha_k = \alpha_{Mp}$ for $k = Mp, \ldots, M(p+1)-1$. By taking the conditional expectation with respect to $\mathcal{G}_p = \{x_0, \ldots, x_{M(p+1)-1}\}$, we have for all $x^* \in X^*$ and p

$$\begin{aligned} E\big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{G}_p\big\} &\leq ||x_{Mp} - x^*||^2 - 2\alpha_{Mp} \sum_{k=Mp}^{M(p+1)-1} E\big\{\big(f_{\omega_k}(x_k) - f_{\omega_k}(x^*)\big) \mid x_k\big\} \\ &\quad + M^2\alpha_{Mp}^2 C_0^2 \\ &\leq ||x_{Mp} - x^*||^2 - \frac{2\alpha_{Mp}}{m} \sum_{k=Mp}^{M(p+1)-1} \big(f(x_k) - f^*\big) + M^2\alpha_{Mp}^2 C_0^2. \end{aligned} \tag{3.6}$$

We now relate $f(x_k)$ and $f(x_{Mp})$ for $k = Mp, \ldots, M(p+1)-1$. We have

$$\begin{aligned} f(x_k) - f^* &= \big(f(x_k) - f(x_{Mp})\big) + \big(f(x_{Mp}) - f^*\big) \\ &\geq \tilde{g}_{Mp}'(x_k - x_{Mp}) + f(x_{Mp}) - f^* \\ &\geq f(x_{Mp}) - f^* - mC_0||x_k - x_{Mp}||, \end{aligned} \tag{3.7}$$

where $\tilde{g}_{Mp}$ is a subgradient of f at x_{Mp} and in the last inequality we use the fact

$$||\tilde{g}_{Mp}|| = \left\|\sum_{i=1}^m \tilde{g}_{i,Mp}\right\| \leq mC_0$$

[cf. Assumption 3.1(b)] with $\tilde{g}_{i,Mp}$ being a subgradient of f_i at x_{Mp}. Furthermore, we have for

all p and $k = Mp, \ldots, M(p+1) - 1$

$$\begin{aligned} ||x_k - x_{Mp}|| &\leq \|x_k - x_{k-1}\| + \|x_{k-1} - x_{Mp}\| \\ &\leq \alpha_{k-1}\|g(\omega_{k-1}, x_{k-1})\| + \|x_{k-1} - x_{Mp}\| \\ &\leq \cdots \\ &\leq \alpha_{Mp} \sum_{l=Mp}^{k-1} ||g(\omega_l, x_l)|| \\ &\leq \big(k - Mp\big)\alpha_{Mp} C_0, \end{aligned} \tag{3.8}$$

which when substituted in Eq. (3.7) yields

$$f(x_k) - f^* \geq f(x_{Mp}) - f^* - \big(k - Mp\big)m\alpha_{Mp}C_0^2.$$

From the preceding relation and Eq. (3.6) we have

$$\begin{aligned} E\big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{G}_{p+1}\big\} \leq ||x_{Mp} - x^*||^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp}) - f^*\big) \\ + 2\alpha_{Mp}^2 C_0^2 \sum_{k=Mp}^{M(p+1)-1} \big(k - Mp\big) + M\alpha_{Mp}^2 C_0^2. \end{aligned} \tag{3.9}$$

Since

$$2\alpha_{Mp}^2 C_0^2 \sum_{k=Mp}^{M(p+1)-1} \big(k - Mp\big) + M\alpha_{Mp}^2 C_0^2 = 2\alpha_{Mp}^2 C_0^2 \sum_{l=1}^{M-1} l + M\alpha_{Mp}^2 C_0^2 = M^2\alpha_{Mp}^2 C_0^2,$$

it follows that for all $x^* \in X^*$ and p

$$E\big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{G}_p\big\} \leq ||x_{Mp} - x^*||^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp}) - f^*\big) + M^2\alpha_{Mp}^2 C_0^2.$$

This relation and the definition of α_k [cf. Eq. (3.5)] yield

$$E\big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{G}_p\big\} \leq ||x_{Mp} - x^*||^2 - \gamma_p\big(2 - \gamma_p\big)\left(\frac{f(x_{Mp}) - f^*}{mC_0}\right)^2.$$

By the supermartingale convergence theorem, we have

$$\sum_{k=0}^{\infty} \gamma_p\big(2 - \gamma_p\big)\left(\frac{f(x_{Mp}) - f^*}{mC_0}\right)^2 < \infty$$

and for each $x^* \in X^*$ the sequence $\{||x_{Mp} - x^*||\}$ is convergent, with probability 1. Because $\gamma_p \in [\underline{\gamma}, \overline{\gamma}] \subset (0, 2)$, it follows that with probability 1

$$\lim_{p\to\infty}\big(f(x_{Mp}) - f^*\big) = 0.$$

Convergence of x_{Mp} to some optimal solution with probability 1 now follows similar to the proof of Prop. 3.2. Since with probability 1

$$||x_k - x_{Mp}|| < M\alpha_{Mp}C, \qquad k = Mp, \ldots, M(p+1) - 1$$

[cf. Eq. (3.8)], it follows that the entire sequence $\{x_k\}$ converges to some optimal solution with probability 1. **Q.E.D.**

Dynamic Stepsize Rule for Unknown f^*

In the case where f^* is not known, we modify the dynamic stepsize (3.5) by replacing f^* with a target level estimate $f_p^{\rm lev}$. Thus the stepsize is

$$\alpha_k = \gamma_p \frac{f(x_{Mp}) - f_p^{\rm lev}}{mMC_0^2}, \quad 0 < \underline{\gamma} \leq \gamma_p \leq \overline{\gamma} < 2, \quad k = Mp, \ldots, M(p+1)-1, \quad p = 0, 1, \ldots. \tag{3.10}$$

To update the target values $f_p^{\rm lev}$, we may use the adjustment procedures described in Section 2.

In the first adjustment procedure, $f_p^{\rm lev}$ is given by

$$f_p^{\rm lev} = \min_{0 \leq j \leq p} f\big(x_{Mj}\big) - \delta_p, \tag{3.11}$$

and δ_p is updated according to

$$\delta_{p+1} = \begin{cases} \delta_p & \text{if } f\big(x_{M(p+1)}\big) \leq f_p^{\rm lev}, \\ \max\big\{\beta\delta_p, \delta\big\} & \text{if } f\big(x_{M(p+1)}\big) > f_p^{\rm lev}, \end{cases} \tag{3.12}$$

where δ and β are fixed positive constants with $\beta < 1$. Thus all the parameters of the stepsize are updated every M iterations. Note that here the parameter ρ of Eq. (2.11) has been set to 1. Our proof relies on this (relatively mild) restriction. Since the stepsize is bounded away from zero, the method behaves similar to the one with a constant stepsize (cf. Prop. 3.1). More precisely, we have the following result.

Proposition 3.4: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method (3.1) and the stepsize rule (3.10) with the adjustment procedure (3.11)–(3.12), we have:

(a) If $f^* = -\infty$, then with probability 1

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then with probability 1

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

Proof: (a) Define the events

$$H_1 = \left\{ \lim_{p \to \infty} \delta_p > \delta \right\}, \qquad H_2 = \left\{ \lim_{p \to \infty} \delta_p = \delta \right\}.$$

Given that H_1 occurred there is an integer R such that $\delta_R > \delta$ and

$$\delta_p = \delta_R, \qquad \forall\ p \geq R.$$

We let R be the smallest integer with the above property and we note that R is a discrete random variable taking nonnegative integer values. In view of Eq. (3.12), we have for all $p \geq R$

$$f\big(x_{M(p+1)}\big) \leq f_p^{\rm lev}.$$

Then from the definition of $f_p^{\rm lev}$ [cf. Eq. (3.11)], the relation $\min_{0\leq j\leq p} f\big(x_{Mj}\big) \leq f\big(x_{Mp}\big)$, and the fact $\delta_p = \delta_R$ for all $p \geq R$, we obtain

$$f\big(x_{M(p+1)}\big) \leq f\big(x_{Mp}\big) - \delta_R, \qquad \forall\ p \geq R.$$

Summation of the above inequalities yields

$$f\big(x_{Mp}\big) \leq f(x_{MR}) - (p-R)\delta_R, \qquad \forall\ p \geq R.$$

Therefore, given that H_1 occurred, we have $\inf_{p\geq 0} f(x_{Mp}) \geq \inf_{p\geq 0} f(x_{Mp}) = -\infty$ with probability 1, i.e.,

$$P\left\{\inf_{p\geq 0} f(x_{Mp}) = -\infty \;\Big|\; H_1\right\} = 1. \tag{3.13}$$

Now assume that H_2 occurred. The event H_2 occurs if and only if, after finitely many iterations, δ_p is decreased to the threshold value δ and remains at that value for all subsequent iterations. Thus H_2 occurs if and only if there is an index S such that

$$\delta_p = \delta, \qquad \forall\ p \geq S. \tag{3.14}$$

Let S be the smallest integer with the above property, and note that we have $H_2 = \cup_{s\geq 0} B_s$, where $B_s = \big\{S = s\big\}$ for all integers $s \geq 0$.

Similar to the proof of Prop. 3.3 [cf. Eq. (3.9)], we have for all $y \in X$ and p

$$\begin{aligned} E\big\{||x_{M(p+1)} - y||^2 \mid \mathcal{G}_p, B_s\big\} &= E\big\{||x_{M(p+1)} - y||^2 \mid \mathcal{G}_p\big\} \\ &\leq ||x_{Mp} - y||^2 - 2\gamma_p \frac{f(x_{Mp}) - f_p^{\rm lev}}{m^2 C_0^2}\big(f(x_{Mp}) - f(y)\big) \\ &\quad + \gamma_p^2 \frac{\big(f(x_{Mp}) - f_p^{\rm lev}\big)^2}{m^2 C_0^2}, \end{aligned} \tag{3.15}$$

where $\mathcal{G}_p = \{x_0, \ldots, x_{Mp-1}\}$. Now, fix an N and let $y_N \in X$ be such that

$$f(y_N) = -N - \delta,$$

where N is a nonnegative integer. Consider a new process $\{\hat{x}_k\}$ defined by

$$\hat{x}_{k+1} = \begin{cases} \mathcal{P}_X\big[\hat{x}_k - \alpha_k g(\omega_k, \hat{x}_k)\big] & \text{if } f(\hat{x}_{Mp}) \geq -N, \\ y_N & \text{otherwise,} \end{cases}$$

for $k = Mp, \ldots, M(p+1)-1$, $p = 0, 1, \ldots$, and $\hat{x}_0 = x_0$. The process $\{\hat{x}_k\}$ is identical to $\{x_k\}$ up to the point when x_{Mp} enters the level set

$$L_N = \{x \in X \mid f(x) < -N\},$$

in which case the process $\{\hat{x}_k\}$ terminates at the point y_N. Therefore, given B_s, the process $\{\hat{x}_{Mp}\}$ satisfies Eq. (3.15) for all $p \geq s$ and $y = y_N$, i.e., we have

$$E\big\{||\hat{x}_{M(p+1)} - y_N||^2 \mid \mathcal{G}_p\big\} \leq ||\hat{x}_{Mp} - y_N||^2 - 2\gamma_p \frac{f(\hat{x}_{Mp}) - f_p^{\rm lev}}{m^2 C_0^2}\big(f(\hat{x}_{Mp}) - f(y_N)\big) + \gamma_p^2 \frac{\big(f(\hat{x}_{Mp}) - f_p^{\rm lev}\big)^2}{m^2 C_0^2},$$

or equivalently

$$E\big\{||\hat{x}_{M(p+1)} - y_N||^2 \mid \mathcal{G}_p\big\} \leq ||\hat{x}_{Mp} - y_N||^2 - z_p,$$

where

$$z_p = \begin{cases} 2\gamma_p \frac{f(\hat{x}_{Mp}) - f_p^{\rm lev}}{m^2 C_0^2}\big(f(\hat{x}_{Mp}) - f(y_N)\big) - \gamma_p^2 \frac{\big(f(\hat{x}_{Mp}) - f_p^{\rm lev}\big)^2}{m^2 C_0^2} & \text{if } \hat{x}_{Mp} \notin L_N, \\ 0 & \text{if } \hat{x}_{Mp} = y_N. \end{cases}$$

By using the definition of $f_p^{\rm lev}$ [cf. Eq. (3.11)] and the fact $\delta_p = \delta$ for all $p \geq s$ [cf. Eq. (3.14)], we have for $p \geq s$ and $\hat{x}_{Mp} \notin L_N$

$$f(y_N) \leq \min_{0 \leq j \leq p} f(\hat{x}_{Mj}) - \delta = f_p^{\rm lev},$$

which when substituted in the preceding relation yields for $p \geq s$ and $\hat{x}_{Mp} \notin L_N$

$$z_p \geq \gamma_p\big(2 - \gamma_p\big)\frac{\big(f(\hat{x}_{Mp}) - f_p^{\rm lev}\big)^2}{m^2 C_0^2} \geq \underline{\gamma}(2 - \overline{\gamma})\frac{\delta^2}{m^2 C_0^2}.$$

The last inequality above follows from the facts $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ and $f(\hat{x}_{Mp}) - f_p^{\rm lev} \geq \delta$ for all p [cf. Eqs. (3.11)–(3.12)]. Hence $z_p \geq 0$ for all k, and by the supermartingale convergence theorem, we obtain $\sum_{p=s}^{\infty} z_p < \infty$ with probability 1. Thus, given B_s and for sufficiently large p, we have $\hat{x}_{Mp} \in L_N$ with probability 1, implying that in the original process

$$P\left\{\inf_{p \geq 0} f(x_{Mp}) \leq -N \;\middle|\; B_s\right\} = 1.$$

By letting $N \to \infty$ in the preceding relation, we obtain

$$P\left\{\inf_{p \geq 0} f(x_{Mp}) = -\infty \;\middle|\; B_s\right\} = 1.$$

Since $H_2 = \cup_{s\geq 0} B_s$, it follows that

$$P\left\{\inf_{p\geq 0} f(x_{Mp}) = -\infty \;\Big|\; H_2\right\} = \sum_{s=0}^{\infty} P\left\{\inf_{p\geq 0} f(x_{Mp}) = -\infty \;\Big|\; B_s\right\} P(B_s) = \sum_{s=0}^{\infty} P(B_s) = 1.$$

Combining Eq. (3.13) with the preceding relation, we have with probability 1

$$\inf_{p\geq 0} f(x_{Mp}) = -\infty,$$

so that $\inf_{k\geq 0} f(x_k) = -\infty$ with probability 1.

(b) Using the proof of part (a), we see that if $f^* > -\infty$, then H_2 occurs with probability 1. Thus, as in part (a), we have $H_2 = \cup_{s\geq 0} B_s$, where $B_s = \{S = s\}$ for all integer $s \geq 0$ and S is as in Eq. (3.14).

Fix an N and let $y_N \in X$ be such that

$$f(y_N) = f^* + \frac{1}{N},$$

where N is a positive integer. Consider the process $\{\hat{x}_k\}$ defined by

$$\hat{x}_{k+1} = \begin{cases} \mathcal{P}_X\big[\hat{x}_k - \alpha_k g(\omega_k, \hat{x}_k)\big] & \text{if } f(\hat{x}_{Mp}) \geq f^* + \delta + \frac{1}{N}, \\ y_N & \text{otherwise,} \end{cases}$$

for $k = Mp, \ldots, M(p+1) - 1$, $p = 0, 1, \ldots$, and $\hat{x}_0 = x_0$. The process $\{\hat{x}_k\}$ is the same as the process $\{x_k\}$ up to the point where x_{Mp} enters the level set

$$L_N = \left\{x \in X \;\Big|\; f(x) < f^* + \delta + \frac{1}{N}\right\},$$

in which case the process $\{\hat{x}_k\}$ terminates at the point y_N. The rest follows similar to the proof of part (a). **Q.E.D.**

The target level f_p^{lev} can also be updated according to the second adjustment procedure discussed in Section 2. In this case, it can be shown that the result of Prop. 2.7 holds with probability 1. We omit the lengthy details.

4. EXPERIMENTAL RESULTS

In this section we report some of the numerical results with a certain type of test problem: the dual of a generalized assignment problem (see Martello and Toth [MaT90], p. 189, Bertsekas [Ber98], p. 362). The problem is to assign m jobs to n machines. If job i is performed at machine

j, it costs a_{ij} and requires p_{ij} time units. Given the total available time t_j at machine j, we want to find the minimum cost assignment of the jobs to the machines. Formally the problem is

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}y_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{n} y_{ij} = 1, \quad i = 1,\dots,m, \\ & \sum_{i=1}^{m} p_{ij}y_{ij} \le t_j, \quad j = 1,\dots,n, \\ & y_{ij} = 0 \text{ or } 1, \quad \text{for all } i,j, \end{aligned}$$

where y_{ij} is the assignment variable, which is equal to 1 if the ith job is assigned to the jth machine and is equal to 0 otherwise. In our experiments we chose n equal to 4 and m equal to the four values 500, 800, 4000, and 7000.

By relaxing the time constraints for the machines, we obtain the dual problem

$$\begin{aligned} \text{maximize} \quad & f(x) = \sum_{i=1}^{m} f_i(x) \\ \text{subject to} \quad & x \ge 0, \end{aligned} \tag{4.1}$$

where

$$f_i(x) = \min_{\sum_{j=1}^{n} y_{ij}=1,\ y_{ij}=0 \text{ or } y_{ij}=1} \sum_{j=1}^{n}(a_{ij} + x_j p_{ij})y_{ij} - \frac{1}{m}\sum_{j=1}^{n} t_j x_j, \qquad i = 1,\dots,m.$$

Since $a_{ij} + x_j p_{ij} \ge 0$ for all i,j, we can easily evaluate $f_i(x)$ for each $x \ge 0$:

$$f_i(x) = a_{ij^*} + x_{j^*}p_{ij^*} - \frac{1}{m}\sum_{j=1}^{n} t_j x_j,$$

where j^* is such that

$$a_{ij^*} + x_{j^*}p_{ij^*} = \min_{1\le j\le n}\{a_{ij} + x_j p_{ij}\}.$$

In the same time, at no additional cost, we obtain a subgradient g of f_i at x:

$$g = (g_1,\dots,g_n)', \qquad g_j = \begin{cases} -\frac{t_j}{m} & \text{if } j \ne j^*, \\ p_{ij^*} - \frac{t_{j^*}}{m} & \text{if } j = j^*. \end{cases}$$

The experiments are divided in two groups, each with a different goal. The first group was designed to compare the performance of the ordinary subgradient method (1.3) and the incremental subgradient method (1.4)–(1.6) for solving the test problem (4.1) when using different stepsize choices, while keeping fixed the order of processing of the components f_i. The second group of experiments was designed to evaluate the incremental method when using different rules for the order of processing the components f_i, while keeping fixed the stepsize choice.

In the first group of experiments the data for the problems (i.e., the matrices $\{a_{ij}\}, \{p_{ij}\}$) were generated randomly according to a uniform distribution over different intervals. The values t_j were calculated according to the formula

$$t_j = \frac{\overline{t}}{n} \sum_{i=1}^{m} p_{ij}, \qquad j = 1, \ldots, n, \tag{4.2}$$

with $\overline{t}$ taking one of the three values 0.5, 0.7, or 0.9. We used two stepsize rules:

(1) A diminishing stepsize that has the form

$$\alpha_{kN} = \cdots = \alpha_{(k+1)N-1} = \frac{D}{k+1}, \qquad \forall\ k \geq 0,$$

where D is some positive constant, and N is some positive integer that represents the number of cycles during which the stepsize is kept at the same value. To guard against an unduly large value of c we implemented an adaptive feature, whereby if within some (heuristically chosen) number S of consecutive iterations the current best cost function value is not improved, then the new iterate x_{k+1} is set equal to the point at which the current best value is attained.

(2) The stepsize rule given by Eq. (2.9) and the path-based procedure. This is essentially the target level method, in which the path bound is not fixed but rather the current value for B is multiplied by a certain factor $\xi \in (0, 1)$ whenever an oscillation is detected (see the remark following Prop. 2.7). The initial value for the path bound was $B = r||x_0 - x_1||$ for some (heuristically chosen) positive constant r.

We report in the following tables the number of iterations required for various methods and parameter choices to achieve a given threshold cost $\tilde{f}$. The notation used in the tables is as follows:

$> k \times 100$ for $k = 1, 2, 3, 4$: means that the value $\tilde{f}$ has been achieved or exceeded after $k \times 100$ iterations, but in less than $(k+1) \times 100$ iterations.

> 500: means that the value $\tilde{f}$ has not been achieved within 500 iterations.

$D/N/S/iter$: gives the values of the parameters D, N, and S for the diminishing stepsize rule, while $iter$ is the number of iterations (or cycles) needed to achieve or exceed $\tilde{f}$.

$r/\xi/\delta_0/iter$: describes the values of the parameters and number of iterations for the target level stepsize rule.

Tables 1 and 2 show the results of applying the ordinary and incremental subgradient methods to problem (4.1) with $n = 4$, $m = 800$, and $\overline{t} = 0.5$ in Eq. (4.2). The optimal value of

the problem is $f^* \approx 1578.47$. The threshold value is $\tilde{f} = 1578$. The tables show when the value $\tilde{f}$ was attained or exceeded.

Ordinary subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.08/2/7/ > 500$	$0.03/0.97/12 \times 10^5/ > 500$
(0,0,0,0)	$0.1/2/7/ > 500$	$0.5/0.98/2 \times 10^4/ > 500$
(0,0,0,0)	$0.07/3/10/ > 500$	$0.5/0.95/3 \times 10^4/ > 500$
(0,0,0,0)	$0.01/10/7/ > 500$	$0.3/0.95/5 \times 10^4/ > 400$
(0,0,0,0)	$0.09/1/7/ > 500$	$0.1/0.9/10^6/ > 200$
(0,0,0,0)	$0.03/5/500/ > 500$	$0.2/0.93/5 \times 10^4/ > 300$
(0,0,0,0)	$0.08/4/7/ > 500$	$0.8/0.97/12 \times 10^3/ > 500$
(0,0,0,0)	$0.09/5/10/ > 500$	$0.03/0.95/10^6/ > 500$
(1.2,1.1,2,1.04)	$0.005/2/5/ > 500$	$0.4/0.975/2 \times 10^4/ > 200$
(1.2,1.1,2,1.04)	$0.009/1/5/ > 500$	$0.5/0.97/4 \times 10^3/ > 50$
$(0.4, 0.2, 1.4, 0.1)$	$0.009/2/5/ > 500$	$0.4/0.8/2700/ > 500$
$(0.4, 0.2, 1.4, 0.1)$	$0.005/5/500/ > 500$	$0.5/0.9/1300/ > 500$

Table 1. $n = 4$, $m = 800$, $f^* \approx 1578.47$, $\tilde{f} = 1578$.

Incremental subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.05/3/500/99$	$3/0.7/5 \times 10^6/97$
(0,0,0,0)	$0.09/2/500/ > 100$	$2/0.6/55 \times 10^5/ > 100$
(0,0,0,0)	$0.1/1/500/99$	$0.7/0.8/55 \times 10^5/ > 100$
(0,0,0,0)	$0.1/1/10/99$	$0.4/0.95/10^7/80$
(0,0,0,0)	$0.05/5/7/ > 100$	$0.3/0.93/10^7/ > 100$
(0,0,0,0)	$0.07/3/10/ > 100$	$0.5/0.9/10^7/ > 200$
(0,0,0,0)	$0.01/7/7/ > 500$	$0.3/0.93/15 \times 10^6/30$
(0,0,0,0)	$0.009/5/7/ > 500$	$2/0.8/5 \times 10^6/ > 100$
(1.2,1.1,2,1.04)	$0.05/1/500/40$	$0.4/0.97/12 \times 10^6/ > 100$
(1.2,1.1,2,1.04)	$0.04/3/500/35$	$0.3/0.975/10^7/27$
(0.4,0.2,1.4,0.1)	$0.07/1/500/48$	$0.4/0.975/12 \times 10^6/100$
(0.4,0.2,1.4,0.1)	$0.048/1/500/39$	$0.5/0.94/12 \times 10^6/ > 100$

Table 2. $n = 4$, $m = 800$, $f^* \approx 1578.47$, $\tilde{f} = 1578$.

Tables 3 and 4 show the results of applying the ordinary and incremental subgradient methods to problem (4.1) with $n = 4$, $m = 4000$, and $\overline{t} = 0.7$ in Eq. (4.2). The optimal value of the problem is $f^* \approx 6832.3$ and the threshold value is $\tilde{f} = 6831.5$. The tables show the number of iterations needed to attain or exceed the value $\tilde{f} = 6831.5$.

Ordinary subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.01/2/7/ > 500$	$1/0.9/5000/58$
(0,0,0,0)	$0.001/5/7/ > 300$	$2/0.99/5500/ > 100$
(0,0,0,0)	$0.0008/5/10/ > 300$	$1.3/0.98/4800/54$
(0,0,0,0)	$0.0005/5/7/ > 200$	$1.5/0.98/2000/88$
(0,0,0,0)	$0.0001/5/10/99$	$0.5/0.8/4000/99$
(0,0,0,0)	$0.0001/2/500/ > 100$	$0.4/0.9/4000/89$
(0,0,0,0)	$0.0001/5/10/ > 200$	$0.5/0.9/3000/88$
(0,0,0,0)	$0.00009/5/500/100$	$0.5/0.95/2000/98$
(0.5,0.9,1.3,0.4)	$0.0005/3/500/ > 100$	$0.5/0.98/2000/95$
(0.5,0.9,1.3,0.4)	$0.0002/7/7/ > 100$	$0.4/0.97/3000/98$
(0.26,0.1,0.18,0.05)	$0.0002/5/7/100$	$0.3/0.98/3000/90$
(0.26,0.1,0.18,0.05)	$0.00005/7/7/30$	$0.095/0.985/10/50$

Table 3. $n = 4$, $m = 4000$, $f^* \approx 6832.3$, $\tilde{f} = 6831.5$.

Incremental subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.005/2/500/46$	$5/0.99/10^6/7$
(0,0,0,0)	$0.007/1/500/37$	$8/0.97/11 \times 10^5/5$
(0,0,0,0)	$0.001/2/500/95$	$2/0.99/7 \times 10^5/ > 100$
(0,0,0,0)	$0.0008/1/500/30$	$0.8/0.4/9 \times 10^5/6$
(0,0,0,0)	$0.0002/2/500/21$	$0.7/0.4/10^6/7$
(0,0,0,0)	$0.0005/2/500/40$	$0.1/0.9/10^6/15$
(0,0,0,0)	$0.0002/2/7/21$	$0.08/0.9/15 \times 10^5/18$
(0,0,0,0)	$0.0003/1/500/21$	$0.25/0.9/2 \times 10^6/20$
(0.5,0.9,1.3,0.4)	$0.001/1/500/40$	$0.07/0.9/10^6/7$
(0.5,0.9,1.3,0.4)	$0.0004/1/500/30$	$0.04/0.9/10^6/26$
(0.26,0.1,0.18,0.05)	$0.00045/1/500/20$	$0.04/0.9/15 \times 10^5/10$
(0.26,0.1,0.18,0.05)	$0.00043/1/7/20$	$0.045/0.91/1.55 \times 10^6/10$

Table 4. $n = 4$, $m = 4000$, $f^* \approx 6832.3$, $\tilde{f} = 6831.5$.

Tables 1 and 2 demonstrate that the incremental subgradient method performs substantially better than the ordinary subgradient method. As m increases, the performance of the incremental method improves as indicated in Tables 3 and 4. The results obtained for other problems that we tested are qualitatively similar and consistently show substantially and often dramatically faster convergence for the incremental method.

We suspected that the random generation of the problem data induced a behavior of the (nonrandomized) incremental method that is similar to the one of the randomized version. Consequently, for the second group of experiments, the coefficients $\{a_{ij}\}$ and $\{p_{ij}\}$ were generated as before and then they were sorted in nonincreasing order, in order to create a sequential dependence among the data. In all runs we used the diminishing stepsize choice (as described earlier) with $S = 500$, while the order of components f_i was changed according to three rules:

(1) *Sorted*. After the data have been randomly generated and sorted, the components are processed in the fixed order $1, 2, \ldots, m$.

(2) *Sorted/Shifted*. After the data have been randomly generated and sorted, they are cyclically shifted by some number K. The components are processed in the fixed order $1, 2, \ldots, m$.

(3) *Random*. The index of the component to be processed is chosen randomly, with each component equally likely to be selected.

To compare fairly the randomized methods with the other methods, we count as an "iteration" the processing of m consecutively and randomly chosen components f_i. In this way, an "iteration" of the randomized method is equally time-consuming as a cycle or "iteration" of any of the nonrandomized methods.

Table 5 below shows the results of applying the incremental subgradient method with order rules (1)–(3) for solving the problem (4.1) with $n = 4$, $m = 800$, and $\overline{t} = 0.9$ in Eq. (4.2). The optimal value is $f^* \approx 1672.44$ and the threshold value is $\tilde{f} = 1672$. The table shows the number of iterations needed to attain or exceed $\tilde{f}$.

Incremental subgradient method / Diminishing stepsize			
Initial point x_0	Sorted order $D/N/iter$	Sorted/Shifted order $D/N/K/iter$	Random order $D/N/iter$
(0,0,0,0)	$0.005/1/ > 500$	$0.007/1/9/ > 500$	0.0095/4/5
(0,0,0,0)	$0.0045/1/ > 500$	$0.0056/1/13/ > 500$	0.08/1/21
(0,0,0,0)	$0.003/2/ > 500$	$0.003/2/7/ > 500$	0.085/1/7
(0,0,0,0)	$0.002/3/ > 500$	$0.002/2/29/ > 500$	0.091/1/17
(0,0,0,0)	$0.001/5/ > 500$	$0.001/6/31/ > 500$	0.066/1/18
(0,0,0,0)	$0.006/1/ > 500$	$0.0053/1/3/ > 500$	0.03/2/18
(0,0,0,0)	$0.007/1/ > 500$	$0.00525/1/11/ > 500$	0.07/1/18
(0,0,0,0)	$0.0009/7/ > 500$	$0.005/1/17/ > 500$	0.054/1/17
(0.2,0.4,0.8,3.6)	$0.001/1/ > 500$	$0.001/1/17/ > 500$	0.01/1/13
(0.2,0.4,0.8,3.6)	$0.0008/3/ > 500$	$0.0008/3/7/ > 500$	0.03/1/8
(0,0.05,0.5,2)	$0.0033/1/ > 400$	$0.0037/1/7/ > 400$	0.033/1/7
(0,0.05,0.5,2)	$0.001/4/ > 500$	$0.0024/2/13/ > 500$	0.017/1/8

Table 5. $n = 4$, $m = 800$, $f^* \approx 1672.44$, $\tilde{f} = 1672$.

Table 6 shows the results of applying the incremental subgradient method with order rules (1)–(3) for solving the problem (4.1) with $n = 4$, $m = 7000$, and $\overline{t} = 0.5$ in Eq. (4.2). The optimal value is $f^* \approx 14601.38$ and the threshold value is $\tilde{f} = 14600$. The tables show when the value $\tilde{f}$ was attained or exceeded.

Tables 5 and 6 show how an unfavorable fixed order can have a dramatic effect on the performance of the incremental subgradient method. Note that shifting the components at the beginning of every cycle did not improve the convergence rate of the method. However, the

randomization of the processing order resulted in fast convergence. The results for the other problems that we tested are qualitatively similar and also demonstrated the superiority of the randomized method.

Incremental subgradient method / Diminishing stepsize			
Initial point x_0	Sorted order $D/N/iter$	Sorted/Shifted order $D/N/K/iter$	Random order $D/N/iter$
(0,0,0,0)	$0.0007/1/ > 500$	$0.0007/1/3/ > 500$	0.047/1/18
(0,0,0,0)	$0.0006/1/ > 500$	$0.0006/1/59/ > 500$	0.009/1/10
(0,0,0,0)	$0.00052/1/ > 500$	$0.00052/1/47/ > 500$	0.008/1/2
(0,0,0,0)	$0.0008/1/ > 500$	$0.0005/1/37/ > 500$	0.023/1/34
(0,0,0,0)	$0.0004/2/ > 500$	$0.0004/2/61/ > 500$	0.0028/1/10
(0,0,0,0)	$0.0003/2/ > 500$	$0.0003/2/53/ > 500$	0.06/1/22
(0,0,0,0)	$0.00025/3/ > 500$	$0.00025/3/11/ > 500$	0.05/1/18
(0,0,0,0)	$0.0009/1/ > 500$	$0.00018/3/79/ > 500$	0.007/1/10
(0,0.1,0.5,2.3)	$0.0005/1/ > 500$	$0.0005/1/79/ > 500$	0.004/1/10
(0,0.1,0.5,2.3)	$0.0003/1/ > 500$	$0.0003/1/51/ > 500$	0.0007/1/18
(0,0.2,0.6,3.4)	$0.0002/1/ > 500$	$0.0002/1/51/ > 500$	0.001/1/10
(0,0.2,0.6,3.4)	$0.0004/1/ > 500$	$0.00007/2/93/ > 500$	0.0006/1/10

Table 6. $n = 4$, $m = 7000$, $f^* \approx 14601.38$, $\tilde{f} = 14600$.

5. CONCLUSIONS

We have proposed several variants of incremental subgradient method, we have analyzed their convergence properties, and we have evaluated them experimentally. The methods that employ the constant and the dynamic stepsize rules are analyzed here for the first time. The subgradient methods of Section 3 are the first incremental methods that use randomization in the context of deterministic nondifferentiable optimization, and their computational performance is particularly interesting. A similar randomization in the context of deterministic differentiable optimization, proposed by Bertsekas and Tsitsiklis [BeT96], p. 143, seems to have a qualitatively different computational performance, as suggested by examples (see Bertsekas [Ber99], p. 113 and p. 616).

Several of the ideas of this paper merit further investigation, some of which will be presented in future publications. In particular, we will discuss in a separate paper variants of the incremental subgradient method involving a momentum term, alternative stepsize rules, the use of ϵ-subgradients, and some other features.

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.

[Ber98] Bertsekas, D. P., Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA., 1998.

[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.

[BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., "Gradient Convergence in Gradient Methods," SIAM J. on Optimization, Vol. 10, No. 3, 2000, pp. 627–642.

[BMN00] Ben-Tal, A., Margalit, T., and Nemirovski, A., "The Ordered Subsets Mirror Descent Optimization Method and its Use for the Positron Emission Tomography Reconstruction," submitted to the Proceedings of the March 2000 Haifa Workshop on Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., Eds., Studies in Computational Mathematics, Elsevier, Amsterdam.

[Brä93] Brännlund, U., "On Relaxation Methods for Nonsmooth Convex Optimization," Doctoral Thesis, Royal Institute of Technology, Stockholm, 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, N.Y., 1985.

[Erm66] Ermoliev, Yu. M., "Methods for Solving Nonlinear Extremal Problems," Kibernetika, Kiev, No. 4, 1966, pp. 1–17.

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

[Erm76] Ermoliev, Yu. M., Stochastic Programming Methods, Nauka, Moscow, 1976.

[Erm83] Ermoliev, Yu. M., "Stochastic Quasigradient Methods and Their Application to System Optimization," Stochastics, Vol. 9, 1983, pp. 1–36.

[Erm88] Ermoliev, Yu. M., "Stochastic Quasigradient Methods," in Numerical Techniques for Stochastic Optimization, Eds., Ermoliev, Yu. M., and Wets, R. J-B., IIASA, Springer-Verlag, 1988, pp. 141–185.

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

[GoK99] Goffin, J. L, 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.

[KaC98] Kaskavelis, C. A., and Caramanis, M. C., "Efficient Lagrangian Relaxation Algorithms for Industry Size Job-Shop Scheduling Problems," IIE Transactions on Scheduling and Logistics, Vol. 30, 1998, pp. 1085–1097.

[Kib79] Kibardin, V. M., "Decomposition into Functions in the Minimization Problem," Automation and Remote Control, Vol. 40, 1980, pp. 1311–1323.

[KiL00] Kiwiel, K. C., and Lindberg, P. O., "Parallel Subgradient Methods for Convex Optimization," submitted to the Proceedings of the March 2000 Haifa Workshop on Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., Eds., Studies in Computational Mathematics, Elsevier, Amsterdam.

[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," Opt. 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," Opt. Methods and Software, Vol. 4, 1994, pp. 103–116.

[MaT90] Martello, S., and Toth, P., Knapsack Problems, J. Wiley, N.Y., 1990.

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

[NBB00] Nedić, A., Bertsekas, D. P., and Borkar, V. S., "Distributed Asynchronous Incremental Subgradient Methods," submitted to the Proceedings of the March 2000 Haifa Workshop on Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., Eds., Studies in Computational Mathematics, Elsevier, Amsterdam.

[NeB99] Nedić, A., and Bertsekas, D. P., "Incremental Subgradient Methods for Nondifferentiable Optimization," Lab. for Info. and Decision Systems Report LIDS-P-2460, Massachusetts Institute of Technology, Cambridge, MA.

[NeB00] Nedić, A., and Bertsekas, D. P., "Convergence Rate of Incremental Subgradient Algorithms," to appear in Stochastic Optimization: Algorithms and Applications, Uryasev, S., and Pardalos, P. M., Eds.

[Pol67] Polyak, B. T., "A General Method of Solving Extremum Problems," Soviet Math. Doklady, Vol. 8, No. 3, 1967, pp. 593–597.

[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.

[Roc70] Rockafellar, R. T., Convex Analysis, Princeton Univ. Press, Princeton, N.J., 1970.

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

[Sol98] Solodov, M. V., "Incremental Gradient Algorithms with Stepsizes Bounded Away From Zero," Computational Opt. and Appl., Vol. 11, 1998, pp. 28–35.

[SoZ98] Solodov, M. V., and Zavriev, S. K., "Error Stability Properties of Generalized Gradient-Type Algorithms," J. Opt. Theory and Appl., Vol. 98, No. 3, 1998, pp. 663–680.

[Tse98] Tseng, P., "An Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule," SIAM J. on Optimization, Vol. 8, No. 2, 1998, 506–531.

[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.

[ZLW99] Zhao, X., Luh, P. B., and Wang, J., "Surrogate Gradient Algorithm for Lagrangian Relaxation," J. Opt. Theory and Appl., Vol. 100, 1999, pp. 699–712.