Stochastic Optimization: Algorithms and Applications (S. Uryasev and P. M. Pardalos, Editors), pp. 263-304

Convergence Rate of Incremental Subgradient Algorithms

Angelia Nedić (angeg@andja.mit.edu)
Massachusetts Institute of Technology, Rm. 35-307,
77 Massachusetts Ave., Cambridge, MA, 02139, USA

Dimitri Bertsekas (dimitrib@mit.edu)
Massachusetts Institute of Technology, Rm. 35-210,
77 Massachusetts Ave., Cambridge, MA, 02139, USA

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 present convergence results and estimates of the convergence rate of a number of variants of incremental subgradient methods, including some that use randomization. The convergence rate estimates are consistent with our computational results, and suggests that the randomized variants perform substantially better than their deterministic counterparts.

Keywords: nondifferentiable optimization, convex programming, incremental subgradient methods, stochastic subgradient methods.

1 Introduction

Throughout this paper, we focus on the problem

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

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

$$\begin{aligned} &\text{maximize} \qquad \sum_{i=1}^{m} c_i' y_i \\ &\text{subject to} \qquad y_i \in Y_i, \quad i = 1, \ldots, m, \qquad \sum_{i=1}^{m} A_i y_i \geq b, \end{aligned} \tag{2}$$

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), where

$$f_i(x) = \max_{y_i \in Y_i} (c_i + A_i' x)' y_i - \beta_i' x, \tag{3}$$

β_i are vectors in $\Re^n$ such that

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

and the set 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) is the subgradient method

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha_k \sum_{i=1}^{m} d_{i,k} \right], \tag{4}$$

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 $X \subset \Re^n$. There is an extensive theory for this method (see e.g. the textbooks by Bertsekas [3], Dem'yanov and Vasil'ev [7], Hiriart-Urruty and Lemaréchal [15], Minoux [23], Polyak [28], Shor [30]).

We consider the incremental subgradient method proposed in Nedić and Bertsekas [25]. It is similar to the standard subgradient method (4), the main difference being 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

$$x_{k+1} = \psi_{m,k}, \tag{5}$$

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

$$\psi_{i,k} = \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k g_{i,k} \right], \quad g_{i,k} \in \partial f_i(\psi_{i-1,k}), \quad i = 1, \ldots, m, \tag{6}$$

starting with

$$\psi_{0,k} = x_k, \tag{7}$$

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. (6) are referred to as the *subiterations* of the kth cycle.

In our paper [25], we proposed a variety of stepsize selection rules and we proved a number of convergence results. We showed that the incremental method exhibits convergence behavior similar to methods that use ϵ-subgradients (see e.g., Bertsekas [3], Correa and Lemaréchal [6], Dem'yanov and Vasil'ev [7], Hiriart-Urruty and Lemaréchal [15], and Polyak [28], p. 144). Our method of analysis is different than the one of the related earlier work of Solodov and Zavriev [31], where an incremental method and some modifications are considered for a diminishing stepsize. Their work addresses a considerably broader class of problems where the component functions may be nonconvex (as well as nondifferentiable), but requires that the set X be compact. We also note that some incremental subgradient methods that are somewhat different than the ones considered here have been proposed by Kaskavelis and Caramanis [16], and Zhao, Luh, and Wang [32], under the name *interleaved subgradient methods*. 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 directions used in the iteration is the sum of the (approximate) subgradients of all the components f_i.

In our paper [25], we also proposed a randomized version of the incremental subgradient method (5)–(7), where the component function f_i in Eq. (6) 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

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

where a random variable ω is uniformly distributed over the index set $\{1, \ldots, m\}$. Our analysis and computational results indicate that the performance of the randomized method is superior to the performance of its deterministic competitors, at least when m is large (see Section 3 of the present paper, and Nedić and Bertsekas [25]).

In this paper we focus on the convergence rate of the incremental subgradient method (5)–(7) and its randomized variant for various stepsize choices including constant, diminishing, and dynamic stepsize rules. In the next section, we discuss the convergence rate of the nonrandomized method and, in Section 3, we consider the convergence rate of randomized versions.

2 Estimates for Convergence Rate of the Incremental Subgradient Method

In this section, we present some convergence results for the incremental subgradient method (5)–(7) for various stepsize rules, and give estimates of their convergence rates. The convergence results have already been given in Nedić and Bertsekas [24] and [25], and are presented without proofs. We give proofs of the convergence rate estimates, which are presented here for the first time.

Throughout this paper, we use the following notation:

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

We also use the defining property of the subgradient of a convex function $h : \Re^n \to \Re$, namely

$$h(x) + g'(z - x) \leq h(z), \qquad \forall\, z \in \Re^n, \quad \forall\, g \in \partial h(x). \tag{8}$$

Since each component f_i is defined as a real-valued convex function over the entire space $\Re^n$, the subdifferential $\partial f_i(x)$ is nonempty and compact for all x and i.

We assume the following:

Assumption 2.1: (Subgradient Boundedness). There exists a scalar C such that

$$||g|| \leq C, \;\; \forall\, g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}), \;\; i = 1, \ldots, m, \;\; k = 0, 1, \ldots. \tag{9}$$

Assumption 2.2: (Existence of an Optimal Solution). The optimal solution set X^* is nonempty.

In many important applications, the set X is compact so that Assumptions 2.1 and 2.2 are satisfied [the set $\cup_{x \in X} \partial f_i(x)$ is compact if X is compact; see e.g. Bertsekas [3], Proposition B.24]. Also, 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, the subgradient boundedness assumption holds for the dual problem (1)–(3), where for each i and for all x the set of subgradients $\partial f_i(x)$ is the convex hull of a finite number of points. Assumption 2.2 can be shown to hold (based on Rockafellar [29], Theorem 9.3) if $\inf_{x \in X} f_i(x)$ is finite for each i, and at least one of the components f_i has bounded level sets.

Even though we have convergence results that only use Assumption 2.1, all estimates of the convergence rate that we give here require both Assumptions 2.1 and 2.2.

In the next proposition, we give a relation, which holds for the incremental method (5)–(7) with any stepsize rule. This relation is frequently used throughout this section.

Proposition 2.1: Let Assumptions 2.1 and 2.2 hold. Then, for a sequence $\{x_k\}$ generated by the incremental subgradient method (5)–(7) with any stepsize α_k, we have for all k

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \leq \Big(dist(x_k, X^*)\Big)^2 - 2\alpha_k\Big(f(x_k) - f^*\Big) + \alpha_k^2 m^2 C^2, \tag{10}$$

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

Proof: By using the definition of the method [cf. Eqs. (5)–(7)], the nonexpansion property of the projection, the boundedness of the subgradients $g_{i,k}$ [cf. Eq. (9)], and the subgradient inequality [cf. Eq. (8)] for each component function f_i, we obtain for any $x^* \in X^*$

$$\begin{aligned}
||\psi_{i,k} - x^*||^2 &= ||\mathcal{P}_X[\psi_{i-1,k} - \alpha_k g_{i,k}] - x^*||^2 \\
&\leq ||\psi_{i-1,k} - \alpha_k g_{i,k} - x^*||^2 \\
&\leq ||\psi_{i-1,k} - x^*||^2 - 2\alpha_k g'_{i,k}(\psi_{i-1,k} - x^*) + \alpha_k^2 C^2 \\
&\leq ||\psi_{i-1,k} - x^*||^2 - 2\alpha_k \Big(f_i(\psi_{i-1,k}) - f_i(x^*) \Big) + \alpha_k^2 C^2, \ \forall\, i, k.
\end{aligned}$$

By adding the above inequalities over i, we have for any $x^* \in X^*$ and all k

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

By strengthening the above inequality, we have for any $x^* \in X^*$ and all k

$$\begin{aligned}
||x_{k+1} - x^*||^2 &\leq ||x_k - x^*||^2 - 2\alpha_k \Big(f(x_k) - f^* \Big) + 2\alpha_k \sum_{i=1}^{m} C ||\psi_{i-1,k} - x_k|| \\
&\quad + \alpha_k^2 m C^2 \\
&\leq ||x_k - x^*||^2 - 2\alpha_k \Big(f(x_k) - f^* \Big) \\
&\quad + \alpha_k^2 C^2 \left(2 \sum_{i=2}^{m} (i-1) + m \right) \\
&= ||x_k - x^*||^2 - 2\alpha_k \Big(f(x_k) - f^* \Big) + \alpha_k^2 m^2 C^2,
\end{aligned}$$

where in the first inequality we use the relation

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

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

$$||\psi_{i,k} - x_k|| \leq \alpha_k i C, \qquad i = 1, \ldots, m, \quad \forall\, k,$$

which follows from Eqs. (5)–(7) and Assumption 2.1. Hence for any $x^* \in X^*$ and all k we have

$$||x_{k+1} - x^*||^2 \le ||x_k - x^*||^2 - 2\alpha_k\Big(f(x_k) - f^*\Big) + \alpha_k^2 m^2 C^2. \qquad (11)$$

By taking the minimum over $x^* \in X^*$ in the above relation, we obtain Eq. (10). **Q.E.D.**

2.1 Constant Stepsize Rule

In this section, we present a convergence result and give estimates of convergence rate for the incremental method (5)–(7) when the constant stepsize rule is employed.

Proposition 2.2: Let Assumption 2.1 hold and let $\{x_k\}$ be a sequence generated by the incremental subgradient method (5)–(7) with the stepsize α_k fixed to some positive constant α.

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

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

(b) If f^* is finite, then

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

where C is as in Assumption 2.1.

Proof: See Proposition 2.1 of Nedić and Bertsekas [25]. **Q.E.D.**

The next proposition gives an estimate of the number K of cycles needed to guarantee that

$$\min_{0\le j\le K} f(x_j) \le f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}.$$

Each cycle consists of the m subiterations indicated in Eq. (6).

Proposition 2.3: Let Assumptions 2.1 and 2.2 hold, and let the sequence $\{x_k\}$ be generated by the incremental subgradient method (5)–(7) with the stepsize α_k fixed to some positive constant α. Then for a positive scalar ϵ we have

$$\min_{0\le k\le K} f(x_k) \le f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}, \qquad (12)$$

where K is given by

$$K = \left\lfloor \frac{\left(dist(x_0, X^*)\right)^2}{\alpha\epsilon} \right\rfloor.$$

Proof: In order to arrive at a contradiction, assume that the relation (12) does not hold, so that for all k with $0 \le k \le K$ we have

$$f(x_k) > f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}.$$

By using this relation in Eq. (10), where α_k is replaced by α, we obtain for all k with $0 \le k \le K$

$$\begin{aligned} \Big(dist(x_{k+1}, X^*)\Big)^2 &\le \Big(dist(x_k, X^*)\Big)^2 - 2\alpha\Big(f(x_k) - f^*\Big) + \alpha^2 m^2 C^2 \\ &\le \Big(dist(x_k, X^*)\Big)^2 - (\alpha^2 m^2 C^2 + \alpha\epsilon) + \alpha^2 m^2 C^2 \\ &= \Big(dist(x_k, X^*)\Big)^2 - \alpha\epsilon. \end{aligned}$$

Summation of the above inequalities over k for $k = 0, \ldots, K$ yields

$$\Big(dist(x_{K+1}, X^*)\Big)^2 \le \Big(dist(x_0, X^*)\Big)^2 - (K+1)\alpha\epsilon,$$

so that

$$\Big(dist(x_0, X^*)\Big)^2 - (K+1)\alpha\epsilon \ge 0,$$

which contradicts the definition of K. **Q.E.D.**

Since every cycle consists of m subiterations, the total number N of subiterations needed for Eq. (12) to hold is given by

$$N = mK = m \left\lfloor \frac{\left(dist(x_0, X^*)\right)^2}{\alpha\epsilon} \right\rfloor. \tag{13}$$

Next, under a strong convexity type assumption, we show that the convergence rate of the incremental subgradient method is linear for a sufficiently small constant stepsize. However, only convergence to a neighborhood of the optimum can be guaranteed.

Proposition 2.4: Let Assumptions 2.1 and 2.2 hold. Also, assume that there exists a positive scalar μ such that

$$f(x) - f^* \ge \mu \Big(dist(x, X^*)\Big)^2, \qquad \forall\, x \in X. \tag{14}$$

Then, for a sequence $\{x_k\}$ generated by the incremental subgradient method (5)–(7) with the stepsize α_k fixed to some positive constant α, where $\alpha \le \frac{1}{2\mu}$, we have

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \le (1 - 2\alpha\mu)^{k+1}\Big(dist(x_0, X^*)\Big)^2 + \frac{\alpha m^2 C^2}{2\mu}. \qquad (15)$$

Proof: By using Eq. (14) in the relation (10) where α_k is replaced by α, we obtain for all k

$$\begin{aligned}\Big(dist(x_{k+1}, X^*)\Big)^2 &\le \Big(dist(x_k, X^*)\Big)^2 - 2\alpha\Big(f(x_k) - f^*\Big) + \alpha^2 m^2 C^2 \\ &\le (1 - 2\alpha\mu)\Big(dist(x_k, X^*)\Big)^2 + \alpha^2 m^2 C^2.\end{aligned}$$

¿From the above relation, by induction, we see that for all k

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \le (1-2\alpha\mu)^{k+1}\Big(dist(x_0, X^*)\Big)^2 + \alpha^2 m^2 C^2 \sum_{j=0}^{k}(1-2\alpha\mu)^j,$$

which combined with the fact $\sum_{j=0}^{k}(1 - 2\alpha\mu)^j \le \frac{1}{2\alpha\mu}$ for all k, yields Eq. (15). **Q.E.D.**

2.2 Diminishing Stepsize Rule

Here we consider the incremental subgradient method (5)–(7) that uses a diminishing stepsize α_k. First, we present convergence results and then we give a convergence rate estimate.

The next result parallels the convergence result for the ordinary subgradient method of Ermoliev [8] (see also Polyak [26]).

Proposition 2.5: Let Assumption 2.1 hold and assume that the stepsize α_k is such that

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

Then, for the sequence $\{x_k\}$ generated by the incremental method (5)–(7), we have

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

Proof: See Proposition 2.2 of Nedić and Bertsekas [25]. **Q.E.D.**

By assuming, in addition, Assumption 2.2 with a compact X^*, Proposition 2.5 can be strengthened. This is the subject of the next proposition, which parallels a well-known convergence result for the ordinary subgradient method (see Shor [30], p. 25, Theorem 2.2). The proposition is similar to a result of Solodov and Zavriev [31] for the incremental subgradient method, which was proved by different methods, and requires that X be a compact set.

Proposition 2.6: Let Assumptions 2.1 and 2.2 hold with compact X^*. Also, assume that the stepsize α_k is such that

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

Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method (5)–(7), we have

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

Proof: See Proposition 2.3 of Nedić and Bertsekas [25]. **Q.E.D.**

Proposition 2.6 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 shown in the following proposition.

Proposition 2.7: Let Assumptions 2.1 and 2.2 hold. Furthermore, assume that the stepsize α_k is such that

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

Then the sequence $\{x_k\}$ converges to an optimal solution.

Proof: Use Eq. (11) and Proposition 1.3 of Correa and Lemaréchal [6]. **Q.E.D.**

We will now estimate the convergence rate of the incremental subgradient method under a strong convexity type assumption on f. For this we need the following result, which is based on the estimates given in Polyak [28], p. 45, Lemma 4.

Lemma 2.1: Let $\{u_k\}$ be a sequence of nonnegative real numbers satisfying

$$u_{k+1} \le \left(1 - \frac{p}{k+1}\right) u_k + \frac{d}{(k+1)^2}, \qquad \forall\, k \ge 0,$$

where p and d are some positive constants. Then

$$\begin{cases} u_{k+1} \le \frac{1}{(k+2)^p}\left(u_0 + \frac{2^p d(2-p)}{1-p}\right) & \text{if } 0 < p < 1, \\ u_{k+1} \le \frac{d\left(1+\ln(k+1)\right)}{k+1} & \text{if } p = 1, \\ u_{k+1} \le \frac{1}{(p-1)(k+2)}\left(d + \frac{(p-1)u_0 - d}{(k+2)^{p-1}}\right) & \text{if } p > 1. \end{cases}$$

Proof: See Proposition 2.5 of Nedić and Bertsekas [24]. **Q.E.D.**

Now, under a strong convexity type assumption, we show that the incremental subgradient method (5)–(7) with the stepsize $\alpha_k = \frac{R}{k+1}$ has a sublinear convergence rate.

Proposition 2.8: Let Assumptions 2.1 and 2.2 hold, and assume that there exists a positive scalar μ such that

$$f(x) - f^* \ge \mu\left(dist(x, X^*)\right)^2, \qquad \forall\, x \in X. \tag{16}$$

Then for the sequence $\{x_k\}$ generated by the incremental subgradient method (5)–(7) with the stepsize of the form $\alpha_k = \frac{R}{k+1}$, where R is some positive scalar, we have

$$\begin{cases} (dist(x_{k+1}, X^*))^2 \le \frac{1}{(k+2)^p}\Big((dist(x_0, X^*))^2 \\ \qquad\qquad + 2^p R^2 m^2 C^2 \frac{2-p}{1-p}\Big) & \text{if } p \in (0,1), \\ (dist(x_{k+1}, X^*))^2 \le \frac{1+\ln(k+1)}{k+1} R^2 m^2 C^2 & \text{if } p = 1, \\ (dist(x_{k+1}, X^*))^2 \le \frac{1}{(p-1)(k+2)}\Bigg(R^2 m^2 C^2 \\ \qquad\qquad + \frac{(p-1)\left(dist(x_0, X^*)\right)^2 - R^2 m^2 C^2}{(k+2)^{p-1}}\Bigg) & \text{if } p > 1, \end{cases}$$

where $p = 2\mu R$.

Proof: By combining Eqs. (10) and (16), we have for all k

$$\left(dist(x_{k+1}, X^*)\right)^2 \le (1 - 2\mu\alpha_k)\left(dist(x_k, X^*)\right)^2 + \alpha_k^2 m^2 C^2.$$

By applying Lemma 2.1 with $u_k = \Big(dist(x_k, X^*)\Big)^2$, $p = 2\mu R$, and $d = R^2 m^2 C^2$, we obtain the desired estimates. **Q.E.D.**

The estimates of Proposition 2.8 are valid even if the inequality (16) holds for all $x \in X$ in a neighborhood of X^*, i.e., for some positive scalar ϵ we have

$$f(x) - f^* \geq \mu \, \Big(dist(x, X^*)\Big)^2, \quad \forall\, x \in X \text{ with } dist(x, X^*) \leq \epsilon. \tag{17}$$

The reason is that the stepsize $\alpha_k = \frac{R}{k+1}$ satisfies the assumption of Proposition 2.6, so that we have $dist(x_k, X^*) \to 0$. Therefore for sufficiently large k, we have $dist(x_k, X^*) \leq \epsilon$. Note also that the relation (17) holds if for all $x \in X$ in some neighborhood of X^* we have

$$f(x) - f^* \geq \mu \, \Big(dist(x, X^*)\Big)^q,$$

with $1 \leq q \leq 2$.

2.3 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

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

with $g_k \in \partial f(x_k)$, suggested by Polyak in [27], (see also discussions in Bertsekas [3], Brännlund [5], and Shor [30]). For the incremental method, we consider a variant of this stepsize of the form

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{m^2 C^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \tag{18}$$

where C is as in Assumption 2.1. For this choice of stepsize we have to be able to calculate the upper bound C, which can be done, for example, when the components f_i are polyhedral.

In the next proposition we give a convergence result for the incremental subgradient method with the dynamic stepsize given by Eq. (18).

Proposition 2.9: Let Assumptions 2.1 and 2.2 hold. Then the sequence $\{x_k\}$ obtained by the incremental subgradient method (5)–(7) with the dynamic stepsize given by Eq. (18) converges to some optimal solution.

Proof: See Proposition 2.5 of Nedić and Bertsekas [25]. **Q.E.D.**

In what follows, we give several estimates of the convergence rate for the incremental subgradient method with the dynamic stepsize of the form (18). In the next proposition, we present an asymptotic estimate for convergence rate of $f(x_k)$, which parallels Theorem 2, p. 142, in Polyak [28] for the ordinary subgradient method, and we estimate the number K of cycles required for

$$\min_{0\le k\le K} f(x_k) - f^* \le \epsilon$$

to hold, where $\epsilon > 0$ is a prescribed error tolerance.

Proposition 2.10: Let Assumptions 2.1 and 2.2 hold. Also, let the sequence $\{x_k\}$ be generated by the incremental subgradient method (5)–(7) with the dynamic stepsize given by Eq. (18).

(a) We have

$$\liminf_{k\to\infty} \sqrt{k}\Big(f(x_k) - f^*\Big) = 0.$$

(b) For a positive scalar ϵ, we have

$$\min_{0\le k\le K} f(x_k) - f^* \le \epsilon, \tag{19}$$

where K is given by

$$K = \left\lfloor \frac{m^2C^2\Big(dist(x_0, X^*)\Big)^2}{\underline{\gamma}(2-\overline{\gamma})\epsilon^2} \right\rfloor, \tag{20}$$

Proof: (a) Assume, to arrive at a contradiction, that

$$\liminf_{k\to\infty} \sqrt{k}(f(x_k) - f^*) = 2\epsilon$$

for some $\epsilon > 0$. Then for k_0 large enough we have $f(x_k) - f^* \ge \frac{\epsilon}{\sqrt{k}}$ for all $k \ge k_0$. Therefore

$$\sum_{k=k_0}^{\infty} \Big(f(x_k) - f^*\Big)^2 \ge \epsilon^2 \sum_{k=k_0}^{\infty} \frac{1}{k} = \infty. \tag{21}$$

On the other hand, by substituting the expression for the stepsize α_k in Eq. (10), we obtain for all $k \geq k_0$

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \leq \Big(dist(x_k, X^*)\Big)^2 - \gamma_k(2-\gamma_k)\frac{\big(f(x_k)-f^*\big)^2}{m^2C^2}, \quad (22)$$

so that

$$\sum_{k=0}^{\infty}\Big(f(x_k)-f^*\Big)^2 < \infty,$$

which contradicts Eq. (21). Hence, we must have

$$\liminf_{k\to\infty} \sqrt{k}(f(x_k)-f^*) = 0.$$

(b) To arrive at a contradiction, assume that Eq. (19) does not hold, so that for all k with $0 \leq k \leq K$ we have

$$f(x_k) - f^* > \epsilon.$$

By substituting the above relation in Eq. (22) and by using the fact $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ for all k, we obtain for all k with $0 \leq k \leq K$

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \leq \Big(dist(x_k, X^*)\Big)^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\epsilon^2}{m^2C^2},$$

Summation of these inequalities over k for $k = 0, \ldots, K$ yields

$$\Big(dist(x_{K+1}, X^*)\Big)^2 \leq \Big(dist(x_0, X^*)\Big)^2 - (K+1)\frac{\underline{\gamma}(2-\overline{\gamma})\epsilon^2}{m^2C^2},$$

so that

$$(K+1)\frac{\underline{\gamma}(2-\overline{\gamma})\epsilon^2}{m^2C^2} \leq \Big(dist(x_0, X^*)\Big)^2,$$

which contradicts the definition of K [cf. Eq. (20)]. **Q.E.D.**

By using Eq. (20), we see that mK is an upper bound on the number N of subiterations required for

$$\min_{0\leq k\leq K} f(x_k) - f^* \leq \epsilon$$

to hold.

The bound on the number K given by Eq. (20), viewed as a function of $\underline{\gamma}$ and $\overline{\gamma}$, is smallest when $\underline{\gamma} = \overline{\gamma} = 1$. For practical use of these bounds, we need some additional information about f or X. For example, if we know an upper bound r on $dist(x_0, X^*)$, then we can estimate K according to Eq. (20).

Under some additional assumptions on f, we can obtain some different types of estimate of the convergence rate for the method (5)–(7) with the dynamic stepsize. In deriving these estimates, we use the following result given by Polyak [28], p. 46, Lemma 6.

Lemma 2.2: Let $\{u_k\}$ be a sequence of positive numbers satisfying for all k

$$u_{k+1} \le u_k - \beta_k u_k^{1+p},$$

where β_k are nonnegative scalars and p is a positive constant. Then

$$u_k \le u_0 \left(1 + p u_0^p \sum_{j=0}^{k-1} \beta_j\right)^{-\frac{1}{p}}, \qquad \forall\ k \ge 0.$$

In particular, if $\beta_k \equiv \beta$, then

$$u_k \le u_0 (1 + p u_0^p \beta k)^{-\frac{1}{p}}, \qquad \forall\ k \ge 0.$$

We have the following.

Proposition 2.11: Let Assumptions 2.1 and 2.2 hold. Also, let the sequence $\{x_k\}$ be generated by the incremental subgradient method (5)–(7) with the dynamic stepsize given by Eq. (18).

(a) If the function f satisfies

$$f(x) - f^* \ge \mu\ dist(x, X^*), \qquad \forall\ x \in X,$$

for some positive scalar μ, then we have

$$dist(x_k, X^*) \le q^k\ dist(x_0, X^*), \qquad \forall\ k \ge 0,$$

where

$$q = \sqrt{1 - \underline{\gamma}(2 - \overline{\gamma})\frac{\mu^2}{m^2 C^2}}.$$

(b) If the function f satisfies

$$f(x) - f^* \ge \mu\ \Big(dist(x, X^*)\Big)^{1+p}, \qquad \forall\ x \in X,$$

for some positive scalars μ and p, then

$$dist(x_k, X^*) \le \frac{dist(x_0, X^*)}{(1+\overline{C}k)^{\frac{1}{2p}}}, \qquad \forall\ k \ge 0,$$

where

$$\overline{C} = p\underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}\Big(dist(x_0, X^*)\Big)^{2p}.$$

Proof: (a) By using the fact $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and the given property of f, from Eq. (10) we obtain for all k

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \le \Big(dist(x_k, X^*)\Big)^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}\Big(dist(x_k, X^*)\Big)^2.$$

Summation of the above inequalities yields the desired estimate.
(b) Similar to part (a), from Eq. (10), the fact $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$, and the given property of f, we obtain for all k

$$\Big(dist(x_{k+1}, X^*)\Big)^2 \le \Big(dist(x_k, X^*)\Big)^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}\Big(dist(x_k, X^*)\Big)^{2(1+p)}.$$

By denoting $u_k = (dist(x_k, X^*))^2$, we can rewrite the preceding inequality as

$$u_{k+1} \le u_k - \beta u_k^{1+p}, \qquad \forall\ k \ge 0,$$

where $\beta = \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}$. Evidently the sequence $\{u_k\}$ satisfies Lemma 2.2. Therefore

$$u_k \le \frac{u_0}{(1+kp\beta u_0^p)^{\frac{1}{p}}}, \qquad \forall\ k \ge 0.$$

By substituting $u_k = (dist(x_k, X^*))^2$ and $\beta = \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}$ in the relation above, after some calculation, we obtain the desired result. **Q.E.D.**

2.4 Dynamic Stepsize Rule for Unknown f^*

In most practical problems the value f^* is not known. In this case, a popular modification of the dynamic stepsize rule for the ordinary subgradient method is

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\text{lev}}}{||g_k||^2}, \quad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2, \quad \forall\ k \ge 0, \tag{23}$$

where $g_k \in \partial f(x_k)$ and the target level $f_k^{\rm lev}$ is an estimate of f^*. This stepsize with a constant target level (i.e., $f_k^{\rm lev} = w$ for some $w > 0$ and all k) was first proposed by Polyak in [27], and further analyzed by Allen, Helgason, Kennington, and Shetty [1], and Kim and Um [18]. The adjustment procedures for the target level $f_k^{\rm lev}$ in Eq. (23) that require knowledge of an underestimate of f^* are presented in Bazaraa and Sherali [2], Kim, Ahn, and Cho [17], Kiwiel [19], [20]. The procedures for $f_k^{\rm lev}$ that do not require any additional information about f^* are considered in Bertsekas [3], Brännlund [5], Goffin and Kiwiel [14], Kiwiel, Larsson, and Lindberg [21], Kulikov and Fazylov [22].

For the incremental subgradient method, we consider a modification of the stepsize in Eq. (23) of the form

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{m^2 C^2}, \quad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2, \quad \forall\, k \ge 0, \tag{24}$$

where $f_k^{\rm lev}$ is a target level estimate of f^*.

We discuss two procedures for updating the target values $f_k^{\rm lev}$ that do not use knowledge of an underestimate of f^*. 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 kth iteration minus a positive amount δ which is adjusted based on the algorithm's progress. The first adjustment procedure originally presented in our paper [25] 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 [5], and Goffin and Kiwiel [14].

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

$$f_k^{\rm lev} = \min_{0 \le j \le k} f(x_j) - \delta_k, \tag{25}$$

where δ_k is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k & \text{if } f(x_{k+1}) < f_k^{\rm lev}, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) \ge f_k^{\rm lev}, \end{cases} \tag{26}$$

where δ_0, δ, β, and ρ are fixed positive constants with $\beta < 1$ and $\rho \ge 1$. This procedure is particularly simple if $\rho = 1$ and $\delta_0 = \delta$, in which case $f_k^{\rm lev} = \min_{0 \le j \le k} f(x_j) - \delta$ for all k.

In the procedure (25)–(26) 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. (24) is bounded away from zero, since we have

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

The effect is that the method behaves similar to the one with a constant stepsize (cf. Proposition 2.2). In particular, we have the following result.

Proposition 2.12: Let Assumption 2.1 hold, and let $\{x_k\}$ be a sequence generated by the incremental target level method (5)–(7) with the dynamic stepsize (24) and the adjustment procedure (25)–(26).

(a) If $f^* = -\infty$, then
$$\inf_{k\geq 0} f(x_k) = f^*.$$

(b) If f^* is finite, then
$$\inf_{k\geq 0} f(x_k) \leq f^* + \delta.$$

Proof: See Proposition 2.6 of Nedić and Bertsekas [25]. **Q.E.D.**

The estimate in the next proposition parallels the one of Proposition 2.3 for the constant stepsize rule. The estimate is based on similar results given by Kiwiel [19], and Kulikov and Fazylov [22], for the ordinary subgradient method.

Proposition 2.13: Let Assumptions 2.1 and 2.2 hold, and let $\{x_k\}$ be a sequence generated by the incremental target level method (5)–(7) with the dynamic stepsize (24) and the adjustment procedure (25)–(26). Then we have

$$\min_{0\leq k\leq K} f(x_k) \leq f^* + \max_{0\leq k\leq K} \delta_k, \tag{27}$$

where K is given by

$$K = \left\lfloor \frac{m^2C^2\Big(dist(x_0, X^*)\Big)^2}{\underline{\gamma}(2-\overline{\gamma})\delta^2} \right\rfloor. \tag{28}$$

Proof: By using the definition of the stepsize [cf. Eq. (24)], from Eq. (10) we obtain for all k

$$\begin{aligned}\Big(dist(x_{k+1},X^*)\Big)^2 \;\le\; & \Big(dist(x_k,X^*)\Big)^2 - 2\gamma_k\frac{f(x_k)-f_k^{\rm lev}}{m^2C^2}\Big(f(x_k)-f^*\Big) \\ & +\gamma_k^2\frac{\Big(f(x_k)-f_k^{\rm lev}\Big)^2}{m^2C^2}. \end{aligned}\tag{29}$$

Now, in order to arrive at a contradiction, assume that Eq. (27) does not hold, so that for all k with $0\le k\le K$ we have

$$f(x_k) > f^* + \max_{0\le j\le K}\delta_j,$$

which implies that for all k with $0\le k\le K$

$$f_k^{\rm lev} = \min_{0\le j\le k} f(x_j) - \delta_k > f^* + \max_{0\le j\le K}\delta_j - \delta_k \ge f^*.$$

By combining the above relation with Eq. (29) and by using the fact $f(x_k)-f_k^{\rm lev}\ge\delta_k$ for all k, we obtain for all k with $0\le k\le K$

$$\begin{aligned}\Big(dist(x_{k+1},X^*)\Big)^2 \;\le\; & \Big(dist(x_k,X^*)\Big)^2 - 2\gamma_k\frac{\Big(f(x_k)-f_k^{\rm lev}\Big)^2}{m^2C^2} \\ & +\gamma_k^2\frac{\Big(f(x_k)-f_k^{\rm lev}\Big)^2}{m^2C^2} \\ \le\; & \Big(dist(x_k,X^*)\Big)^2 - \gamma_k(2-\gamma_k)\frac{\delta_k^2}{m^2C^2} \\ \le\; & \Big(dist(x_k,X^*)\Big)^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{m^2C^2}, \end{aligned}$$

where the last inequality above follows from the facts $\gamma_k\in[\underline{\gamma},\overline{\gamma}]$ and $\delta_k\ge\delta$ for all k. Summation of the above inequalities over k for $k=0,1,\ldots,K$ yields

$$\Big(dist(x_{K+1},X^*)\Big)^2 \le \Big(dist(x_0,X^*)\Big)^2 - (K+1)\frac{\underline{\gamma}(2-\overline{\gamma})\delta^2}{m^2C^2},$$

so that

$$(K+1)\frac{\underline{\gamma}(2-\overline{\gamma})\delta^2}{m^2C^2} \le \Big(dist(x_0,X^*)\Big)^2,$$

which contradicts the definition of K [see Eq. (28)]. **Q.E.D.**

If $\rho = 1$ and $\delta_0 = \delta$ in Eq. (26), then we have $\delta_k = \delta$ for all k, so that the estimate (27) holds with δ replacing $\max_{0\le k\le K}\delta_k$.

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 a sufficient descent.

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 $f_k^{\rm rec}$ is 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. (5)–(7).

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

We can group the iterates k into sets $I_l = \{k(l), k(l)+1, \ldots, k(l+1)-1\}$ for $l \ge 0$, so that for all k in a given set I_l the incremental target level method aims at the same target level $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$. The target level is updated only if sufficient descent or oscillation is detected (Step 2 or Step 3, respectively). From Eqs. (5)–(7) it can be seen 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 $k \in I_l$. Whenever σ_k exceeds the prescribed upper bound B on the path length, the parameter δ_l is halved.

In the next proposition we present a convergence result for the incremental target level algorithm, which parallels the result given by Goffin and Kiwiel [14] for the ordinary subgradient method.

Proposition 2.14: Let Assumption 2.1 hold. Then, for a sequence $\{x_k\}$ generated by the incremental target level algorithm, we have

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

Proof: See Proposition 2.7 of Nedić and Bertsekas [25]. **Q.E.D.**

For the incremental target level algorithm, we can estimate the number K of cycles required for

$$\min_{0 \leq k \leq K} f(x_k) - f^* \leq \delta_0$$

to hold, as shown in the next proposition.

Proposition 2.15: Let Assumptions 2.1 and 2.2 hold. Then, for a sequence $\{x_k\}$ generated by the incremental target level algorithm, we have

$$\min_{0 \leq k \leq K} f(x_k) - f^* \leq \delta_0,$$

where K is the largest positive integer such that

$$\sum_{k=0}^{K-1} \gamma_k (2 - \gamma_k) \delta_k^2 \leq m^2 C^2 \left(dist(x_0, X^*) \right)^2,$$

and $\delta_k = \delta_l$ for all $k \in I_l$ and all l.

Proof: We use the fact $\delta_l \leq \delta_0$ for all l and we follow the line of analysis of the proof of Proposition 2.13. **Q.E.D.**

3 An Incremental Subgradient Method with Randomization

All the results of Section 2 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. Namely, 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 of processing the components can significantly affect the rate of the convergence of the method (see the experimental results in Nedić and Bertsekas [25]). Unfortunately, to determine an 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 [4], 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 in analogy with the analysis for stochastic subgradient methods (see e.g., Ermoliev [9], [10], [11], Ermoliev and Wets [12], Polyak [28] p. 159).

The formal description of the randomized method is as follows

$$x_{k+1} = \mathcal{P}_X[x_k - \alpha_k g(\omega_k, x_k)], \tag{30}$$

where $x_0 \in X$ is a given point, 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. The stepsize α_k may be deterministic or dependent on x_k.

Throughout this section we assume the following.

Assumption 3.1:

(a) The sequence $\{\omega_k\}$ is a sequence of independent random variables each of which is 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 $\{g(\omega_k, x_k) \mid k = 0, 1, \ldots\}$ used by the randomized method (30) is bounded, i.e., there exists a positive constant C such that for all k

$$||g(\omega_k, x_k)|| \le C,$$

with probability 1.

Note that if the set X is compact or the components f_i are polyhedral, then Assumption 3.1(b) is satisfied.

We rely on Assumption 3.1 for all convergence results of this section. However, for the estimates of convergence rate, in addition, we use the following.

Assumption 3.2: The set X^* of optimal solutions is nonempty.

In what follows we present convergence results and estimates of the convergence rate for the randomized method (30) with the constant, the diminishing and the dynamic stepsize rules. Some of these proofs rely on the supermartingale convergence theorem, as stated in Bertsekas and Tsitsiklis [4], p. 148.

Theorem 3.1: (*Supermartingale Convergence Theorem*) Let Y_k, X_k, and Z_k, $k = 0, 1, \ldots$, be three sequences of random variables and let $\mathcal{F}_k$, $k = 0, 1, \ldots$, 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, X_k and Z_k are nonnegative, and are functions of the random variables in $\mathcal{F}_k$.

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

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

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

3.1 Constant Stepsize Rule

In this section, we consider the randomized method (30) with a constant stepsize. We start with a result, which we use here and in the next section, where we analyze a diminishing stepsize.

Proposition 3.1: Let Assumptions 3.1 and 3.2 hold. Then, for a sequence $\{x_k\}$ generated by the randomized method (30) with a deterministic stepsize α_k, we have for all k

$$\begin{aligned} E\Big\{(dist(x_{k+1}, X^*))^2\Big\} \leq E\Big\{(dist(x_k, X^*))^2\Big\} \\ -\frac{2\alpha_k}{m} E\Big\{f(x_k) - f^*\Big\} + \alpha_k^2 C^2, \end{aligned} \tag{31}$$

where C is as in Assumption 3.1(b).

Proof: By using Eq. (30), the nonexpansion property of the projection, and the boundedness of the subgradients $g(\omega_k, x_k)$, we have for any $x^* \in X^*$ and all k

$$\begin{aligned} ||x_{k+1} - x^*||^2 &= ||\mathcal{P}_X[x_k - \alpha_k g(\omega_k, x_k)] - x^*||^2 \\ &\leq ||x_k - \alpha_k g(\omega_k, x_k) - x^*||^2 \\ &\leq ||x_k - x^*||^2 + \alpha_k^2 C^2 - 2\alpha_k g(\omega_k, x_k)'(x_k - x^*). \end{aligned}$$

By taking the expectation conditioned on $\mathcal{F}_k = \{x_0, x_1, \ldots, x_k\}$ in the above inequality, we obtain for any $x^* \in X^*$ and all k

$$\begin{aligned} E\Big\{||x_{k+1} - x^*||^2 \mid \mathcal{F}_k\Big\} &\leq ||x_k - x^*||^2 + \alpha_k^2 C^2 \\ &\quad -2\alpha_k E\Big\{g(\omega_k, x_k)'(x_k - x^*) \mid x_k\Big\} \\ &\leq ||x_k - x^*||^2 + \alpha_k^2 C^2 \\ &\quad -\frac{2\alpha_k}{m}\Big(f(x_k) - f^*\Big), \end{aligned} \tag{32}$$

where in the last inequality we use the fact

$$\begin{aligned} E\Big\{g(\omega_k, x_k)'(x_k - x^*) \mid \mathcal{F}_k\Big\} &= E\Big\{g(\omega_k, x_k)'(x_k - x^*) \mid x_k\Big\} \\ &\geq \frac{1}{m}\sum_{i=1}^{m}\Big(f_i(x_k) - f_i(x^*)\Big) \\ &= \frac{1}{m}\Big(f(x_k) - f^*\Big), \end{aligned}$$

which follows from the properties of ω_k and the convexity of each f_i. Finally, by taking the minimum over $x^* \in X^*$ of both sides in the relation (32) and by using the fact

$$\min_{x^* \in X^*} E\Big\{||x_{k+1} - x^*||^2 \mid \mathcal{F}_k\Big\} \geq E\Big\{\Big(dist(x_{k+1}, X^*)\Big)^2 \mid \mathcal{F}_k\Big\},$$

we obtain for all k

$$\begin{aligned} E\Big\{\Big(dist(x_{k+1}, X^*)\Big)^2 \mid \mathcal{F}_k\Big\} \leq& \Big(dist(x_k, X^*)\Big)^2 \\ &-\frac{2\alpha_k}{m}\Big(f(x_k) - f^*\Big) + \alpha_k^2 C^2, \end{aligned} \tag{33}$$

which, after taking the total expectation, yields Eq. (31). **Q.E.D.**

In the next proposition, we give a convergence result for the randomized method when a constant stepsize is employed.

Proposition 3.2: Let Assumption 3.1 hold and let the stepsize α_k be fixed to a positive constant α. Also, let a sequence $\{x_k\}$ be generated by the randomized method (30).

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

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

(b) If f^* is finite, then with probability 1 we have

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

Proof: See Proposition 3.1 of Nedić and Bertsekas [25]. **Q.E.D.**

The estimate in part (b) is sharp. For example, take $f_i(x) = C|x|$ for all $x \in \Re$ and $i = 1, \ldots, m$. For any α, choose the initial point $x_0 = \frac{\alpha C}{2}$. In this case, it can be seen that the iterates x_k generated by the method (30) take the values $\frac{\alpha C}{2}$ or $-\frac{\alpha C}{2}$, so that for all k

$$f(x_k) = \frac{\alpha m C^2}{2}.$$

Furthermore, by comparing the error bounds of Propositions 2.2 and 3.2, we see that for the same value of the stepsize α the error bound is smaller by a factor of m for the randomized order method of this section than that for the fixed order method of Section 2.

The estimate of the next proposition parallels that of Proposition 2.3 for the nonrandomized incremental subgradient method.

Proposition 3.3: Let Assumptions 3.1 and 3.2 hold, and let the sequence $\{x_k\}$ be generated by the randomized method (30) with the stepsize α_k fixed to some positive constant α. Then, for a positive scalar ϵ, we have with probability 1

$$\min_{0\leq k\leq N} f(x_k) \leq f^* + \frac{\alpha m C^2 + \epsilon}{2}, \tag{34}$$

where N is a random variable with

$$E\Big\{N\Big\} \leq \frac{m\Big(dist(x_0, X^*)\Big)^2}{\alpha\epsilon}. \tag{35}$$

Proof: Define a new process $\{\tilde{x}_k\}$ as follows

$$\tilde{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } f(x_k) \geq f^* + \frac{\alpha m C^2+\epsilon}{2}, \\ \tilde{y} & \text{otherwise,} \end{cases}$$

where $\tilde{x}(0) = x(0)$ and $\tilde{y}$ is some fixed vector in X^*. The process $\{\tilde{x}_k\}$ is identical to $\{x_k\}$, except that once x_k enters the level set

$$L = \left\{ x \in X \mid f(x) < f^* + \frac{\alpha m C^2 + \epsilon}{2} \right\}$$

the process $\{\tilde{x}_k\}$ terminates at the point $\tilde{y}$. Then for the process $\{\tilde{x}_k\}$ we have for all k [cf. Eq. (33) with $\alpha_k = \alpha$]

$$\begin{aligned} E\Big\{(dist(\tilde{x}_{k+1}, X^*))^2 \mid \mathcal{F}_k\Big\} &\le \Big(dist(\tilde{x}_k, X^*)\Big)^2 - \frac{2\alpha}{m}\Big(f(\tilde{x}_k) - f^*\Big) + \alpha^2 C^2 \\ &= \Big(dist(\tilde{x}_k, X^*)\Big)^2 - z_k, \end{aligned} \tag{36}$$

where $\mathcal{F}_k = \{x_0, \ldots, x_k\}$ and

$$z_k = \begin{cases} \frac{2\alpha}{m}\Big(f(\tilde{x}_k) - f^*\Big) - \alpha^2 C^2 & \text{if } \tilde{x}_k \notin L, \\ 0 & \text{otherwise.} \end{cases}$$

In the case where $\tilde{x}_k \notin L$, we have

$$\begin{aligned} z_k &\ge \frac{2\alpha}{m}\left(f^* + \frac{\alpha m C^2 + \epsilon}{2} - f^*\right) - \alpha^2 C^2 \\ &= \frac{\alpha\epsilon}{m}. \end{aligned} \tag{37}$$

By the supermartingale convergence theorem, from Eq. (36) we have

$$\sum_{k=0}^{\infty} z_k < \infty$$

with probability 1, so that $z_k = 0$ for all $k \ge N$, where N is a random variable. Hence $\tilde{x}_N \in L$ with probability 1, implying that in the original process we have

$$\min_{0 \le k \le N} f(x_k) \le f^* + \frac{\alpha m C^2 + \epsilon}{2}$$

with probability 1. Furthermore, by taking the total expectation in Eq. (36), we obtain for all k

$$\begin{aligned} E\Big\{(dist(\tilde{x}_{k+1}, X^*))^2\Big\} &\le E\Big\{(dist(\tilde{x}_k, X^*))^2\Big\} - E\Big\{z_k\Big\} \\ &\le \Big(dist(x_0, X^*)\Big)^2 - E\left\{\sum_{j=0}^{k} z_j\right\}, \end{aligned}$$

where in the last inequality we use the facts $\tilde{x}_0 = x_0$ and

$$E\Big\{(dist(x_0, X^*))^2\Big\} = \Big(dist(x_0, X^*)\Big)^2.$$

Therefore

$$\Big(dist(x_0, X^*)\Big)^2 \geq E\left\{\sum_{k=0}^{\infty} z_k\right\} = E\left\{\sum_{k=0}^{N-1} z_k\right\} \geq E\left\{\frac{N\alpha\epsilon}{m}\right\} = \frac{\alpha\epsilon}{m}E\Big\{N\Big\},$$

where the last inequality above follows from Eq. (37). **Q.E.D.**

In Section 2, for the nonrandomized incremental method implemented with a constant stepsize α, we showed that

$$\min_{0\leq k\leq K} f(x_k) \leq f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}$$

holds after N iterations, where x_k is the iteration at the end of kth cycle [see Eqs. (5)–(7)] and

$$N = m\left\lfloor \frac{\big(dist(x_0, X^*)\big)^2}{\alpha\epsilon} \right\rfloor$$

[cf. Eq. (13)]. If in the randomized method (30) we use the same stepsize α, then, according to Proposition 3.3, we have with probability 1

$$\min_{0\leq k\leq N} f(x_k) \leq f^* + \frac{\alpha m C^2 + \epsilon}{2},$$

where the expected value of N satisfies

$$E\Big\{N\Big\} \leq \frac{m\Big(dist(x_0, X^*)\Big)^2}{\alpha\epsilon}.$$

Thus, for the same value of ϵ, the bound on the number of iterations for the fixed order method is the same as the bound on the expected number of iterations for the randomized order method. However, the error term $\alpha m^2 C^2$ in the fixed order method is m times larger than the corresponding error term in the randomized order method. Similarly, if we choose the stepsize α in the randomized method to achieve the same error level as in the nonrandomized method, then the corresponding expected number of iterations becomes m times smaller.

Now we give a different estimate of the convergence rate for the randomized method (30) with the constant stepsize rule, and with f satisfying a strong convexity type assumption. The result parallels that of Proposition 2.4.

Proposition 3.4: Let Assumptions 3.1 and 3.2 hold. Also, assume that for some positive scalar μ, with probability 1, we have

$$f(x) - f^* \geq \mu \left(dist(x, X^*)\right)^2, \qquad \forall\, x \in X. \tag{38}$$

Then, for a sequence $\{x_k\}$ generated by the method (30) with a stepsize α_k fixed to some positive scalar α, we have for all k

$$E\Big\{(dist(x_{k+1}, X^*))^2\Big\} \leq \left(1 - \frac{2\alpha\mu}{m}\right)^{k+1} \left(dist(x_0, X^*)\right)^2 + \frac{\alpha m C^2}{2\mu}.$$

Proof: By replacing α_k with α in Eq. (31), we obtain for all k

$$\begin{aligned} E\Big\{(dist(x_{k+1}, X^*))^2\Big\} &\leq E\Big\{(dist(x_k, X^*))^2\Big\} \\ &\qquad -\frac{2\alpha}{m} E\Big\{f(x_k) - f^*\Big\} + \alpha^2 C^2 \\ &\leq \left(1 - \frac{2\alpha\mu}{m}\right) E\Big\{(dist(x_k, X^*))^2\Big\} + \alpha^2 C^2, \end{aligned}$$

where in the last inequality we use the fact

$$E\Big\{f(x_k) - f^*\Big\} \geq \mu\, E\Big\{(dist(x_k, X^*))^2\Big\},$$

which follows from Eq. (38). By induction, we see that for all k

$$\begin{aligned} E\Big\{(dist(x_{k+1}, X^*))^2\Big\} \leq \left(1 - \frac{2\alpha\mu}{m}\right)^{k+1} & E\Big\{(dist(x_0, X^*))^2\Big\} \\ & + C^2\alpha^2 \sum_{j=0}^{k} \left(1 - \frac{2\alpha\mu}{m}\right)^j. \end{aligned}$$

By using the fact

$$E\Big\{(dist(x_0, X^*))^2\Big\} = \left(dist(x_0, X^*)\right)^2$$

and the estimate

$$\sum_{j=0}^{k} \left(1 - \frac{2\alpha\mu}{m}\right)^j \leq \sum_{j=0}^{\infty} \left(1 - \frac{2\alpha\mu}{m}\right)^j = \frac{m}{2\alpha\mu}$$

in the above relation, we obtain the desired result. **Q.E.D.**

Let us compare the estimates of Propositions 2.4 and 3.4. In both propositions, the error bound consists of two terms: the exponentially decreasing term and the asymptotic term. For the same value of the stepsize α, the asymptotic term in the error bound of Proposition 2.4 is m times larger than that of Proposition 3.4. However, if in Proposition 3.4 the stepsize α is replaced by $m\alpha$, then the asymptotic terms and the exponentially decreasing terms in the error bounds of Propositions 2.4 and 3.4, respectively, are the same. The main difference is that in Proposition 2.4, k represents the number of cycles, while in Proposition 3.4, k represents the number of iterations, so that for the same error level the fixed order method requires a number of iterations that is m times larger than that of the randomized order method.

3.2 Diminishing Stepsize Rule

In this section, we consider the randomized method (30) with a diminishing stepsize. First, we present a convergence result.

Proposition 3.5: Let Assumptions 3.1 and 3.2 hold, the stepsize α_k be such that

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

Then, with probability 1, the sequence $\{x_k\}$ generated by the randomized method (30) converges to some optimal solution.

Proof: See Proposition 3.2 of Nedić and Bertsekas [25]. **Q.E.D.**

Next, we give a convergence rate estimate of the randomized method (30) when f satisfies a strong convexity type assumption, and the stepsize is $\alpha_k = \frac{R}{k+1}$ for some positive scalar R.

Proposition 3.6: Let Assumptions 3.1 and 3.2 hold. Also, assume that for some positive scalar μ, with probability 1, we have

$$f(x) - f^* \geq \mu \left(dist(x, X^*)\right)^2, \qquad \forall\, x \in X. \tag{39}$$

Then, for a sequence $\{x_k\}$ generated by the randomized method (30) with the stepsize $\alpha_k = \frac{R}{k+1}$ for some scalar $R > 0$, we have

$$\begin{cases} E\Big\{\big(dist(x_{k+1}, X^*)\big)^2\Big\} \leq \frac{1}{(k+2)^p}\Big(\big(dist(x_0, X^*)\big)^2 \\ \qquad\qquad +2^p C^2 R^2 \frac{2-p}{1-p}\Big) & \text{if } 0 < p < 1, \\ E\Big\{\big(dist(x_{k+1}, X^*)\big)^2\Big\} \leq \frac{C^2R^2\big(1+\ln(k+1)\big)}{k+1} & \text{if } p = 1, \\ E\Big\{\big(dist(x_{k+1}, X^*)\big)^2\Big\} \leq \frac{1}{(p-1)(k+2)}\Bigg(C^2R^2 \\ \qquad\qquad + \frac{(p-1)\big(dist(x_0,X^*)\big)^2 - C^2R^2}{(p-1)(k+2)^{p-1}}\Bigg) & \text{if } p > 1, \end{cases}$$

where $p = \frac{2\mu R}{m}$.

Proof: From Eq. (31) we have for all k

$$\begin{aligned} E\Big\{\big(dist(x_{k+1}, X^*)\big)^2\Big\} &\leq E\Big\{\big(dist(x_k, X^*)\big)^2\Big\} \\ &\quad -\frac{2\alpha_k}{m}E\Big\{f(x_k) - f^*\Big\} + \alpha_k^2 C^2 \\ &\leq \left(1 - \frac{2\alpha_k \mu}{m}\right) E\Big\{\big(dist(x_k, X^*)\big)^2\Big\} + \alpha_k^2 C^2, \end{aligned}$$

where in the last inequality we use the fact

$$E\Big\{f(x_k) - f^*\Big\} \geq \mu\ E\Big\{\big(dist(x_k, X^*)\big)^2\Big\},$$

which follows from Eq. (39). By applying Lemma 2.1 with $u_k = E\Big\{\big(dist(x_k, X^*)\big)^2\Big\}$, $p = \frac{2\mu R}{m}$, and $d = C^2R^2$, and by using

$$u_0 = E\Big\{\big(dist(x_0, X^*)\big)^2\Big\} = \Big(dist(x_0, X^*)\Big)^2,$$

we obtain the desired estimates. **Q.E.D.**

3.3 Dynamic Stepsize Rule for Known f^*

In this section, we present convergence results and the corresponding estimates of the convergence for the randomized method with the dynamic stepsize rule when the optimal function value f^* is known.

A possible version of the dynamic stepsize rule for the method (30) has the form

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

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

$$\alpha_k = \gamma_p \frac{f_p - f^*}{mMC^2}, \;\; 0 < \underline{\gamma} \leq \gamma_p \leq \overline{\gamma} < 2, \;\; k = Mp, \ldots, M(p{+}1){-}1, \quad (40)$$

where $\{\gamma_p\}$, f_p is defined as the value of f at the iterate x_{Mp}. Thus the iterations x_k for $k = Mp, \ldots, M(p{+}1){-}1$ can be viewed as subiterations of the pth cycle. During a cycle, the stepsize is fixed and is updated at the end of a cycle. 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.

We start with a preliminary result, which will also be useful in the case where f^* in Eq. (40) is replaced by an estimate of f^*.

Proposition 3.7: Let Assumptions 3.1 and 3.2 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method (30) with the dynamic stepsize (40), we have for all p

$$E\Big\{(dist(x_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} \leq \Big(dist(x_{Mp}, X^*)\Big)^2 - \frac{2M\alpha_{Mp}}{m}\Big(f(x_{Mp}) - f^*\Big) + M^2C^2\alpha_{Mp}^2 \,, \quad (41)$$

where $\mathcal{F}_p = \Big\{x_0, x_1, \ldots, x_{M(p+1)-1}\Big\}$.

Proof: By using the nonexpansion property of the projection and Assumption 3.1(b), from Eq. (30) it follows that for any $x^* \in X^*$ and all k we have

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - 2\alpha_k g(\omega_k, x_k)'(x_k - x^*) + \alpha_k^2 C^2.$$

By summation of the preceding relations over k for $k = Mp, \ldots, M(p+1) - 1$, where $p \geq 0$ is an integer, (i.e. the summation is over the M iterations of a cycle) and by using the fact

$$\alpha_k = \alpha_{Mp}, \qquad \forall\; k = Mp, \ldots, M(p+1) - 1,$$

we obtain

$$||x_{M(p+1)} - x^*||^2 \le ||x_{Mp} - x^*||^2 \\ -2\alpha_{Mp} \sum_{j=Mp}^{M(p+1)-1} g(\omega_j, x_j)'(x_j - x^*) + MC^2\alpha_{Mp}^2.$$

Taking the expectation conditioned on $\mathcal{F}_p$ yields

$$\begin{aligned} E\Big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{F}_p\Big\} &\le ||x_{Mp} - x^*||^2 + MC^2\alpha_{Mp}^2 \\ &\quad -2\alpha_{Mp} \sum_{j=Mp}^{M(p+1)-1} E\Big\{g(\omega_j, x_j)'(x_j - x^*) \mid x_j\Big\} \\ &\le ||x_{Mp} - x^*||^2 + MC^2\alpha_{Mp}^2 \\ &\quad -\frac{2\alpha_{Mp}}{m} \sum_{j=Mp}^{M(p+1)-1} \Big(f(x_j) - f^*\Big), \end{aligned} \tag{42}$$

where in the last inequality we use the fact

$$E\Big\{g(\omega_j, x_j)'(x_j - x^*) \mid x_j\Big\} \ge \frac{1}{m}\sum_{i=1}^{m}\Big(f_i(x_j) - f_i(x^*)\Big) = \frac{1}{m}\Big(f(x_j) - f^*\Big), \tag{43}$$

which follows from the properties of ω_j [cf. Assumption 3.1(a)] and the convexity of each f_i. Now we need to relate $f(x_j)$ and $f(x_{Mp})$ for $j = Mp, \ldots, M(p+1)-1$. We have

$$\begin{aligned} f(x_j) - f^* &= \Big(f(x_j) - f(x_{Mp})\Big) + \Big(f(x_{Mp}) - f^*\Big) \\ &\ge \tilde{g}_{Mp}'(x_j - x_{Mp}) + f(x_{Mp}) - f^* \\ &\ge f(x_{Mp}) - f^* - ||\tilde{g}_{Mp}|| \cdot ||x_j - x_{Mp}|| \\ &\ge f(x_{Mp}) - f^* - \Big(j - Mp\Big)mC^2\alpha_{Mp}, \end{aligned} \tag{44}$$

where $\tilde{g}_{Mp}$ is a subgradient of f at x_{Mp}, and the last inequality follows from the fact

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

[cf. Assumption 3.1(b)] and the relation

$$||x_j - x_{Mp}|| \le \alpha_{Mp} \sum_{l=Mp}^{j-1} ||g(\omega_l, x_l)|| \le (j - Mp)C\alpha_{Mp},$$

which follows from Eq. (30) and holds for $j = Mp, \ldots, M(p+1) - 1$. By combining Eqs. (43) and (44) with the relation (42), we obtain

$$\begin{aligned} E\Big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{F}_p\Big\} &\leq ||x_{Mp} - x^*||^2 - \frac{2M\alpha_{Mp}}{m}\Big(f(x_{Mp}) - f^*\Big) \\ &\quad + 2C^2\alpha_{Mp}^2 \sum_{j=Mp}^{M(p+1)-1} (j - Mp) + MC^2\alpha_{Mp}^2 \\ &\leq ||x_{Mp} - x^*||^2 - \frac{2M\alpha_{Mp}}{m}\Big(f(x_{Mp}) - f^*\Big) \\ &\quad + M^2C^2\alpha_{Mp}^2, \end{aligned} \tag{45}$$

where in the last inequality we use the fact

$$\begin{aligned} 2C^2\alpha_{Mp}^2 &\sum_{j=Mp}^{M(p+1)-1} (j - Mp) + MC^2\alpha_{Mp}^2 \\ &= 2C^2\alpha_{Mp}^2 \sum_{l=1}^{M-1} l + MC^2\alpha_{Mp}^2 = M^2C^2\alpha_{Mp}^2. \end{aligned}$$

Finally, by taking the minimum over $x^* \in X^*$ in Eq. (45) and by using the fact

$$\min_{x^* \in X^*} E\Big\{||x_{M(p+1)} - x^*||^2 \mid \mathcal{F}_p\Big\} \geq E\Big\{(dist(x_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\},$$

we obtain the relation (41). **Q.E.D.**

Here is a convergence result for the randomized method when the dynamic stepsize given by Eq. (40) is used.

Proposition 3.8: Let Assumptions 3.1 and 3.2 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method (30) with the dynamic stepsize (40), with probability 1, we have

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

Proof: See Proposition 3.3 of Nedić and Bertsekas [25]. **Q.E.D.**

The next proposition parallels Proposition 2.10.

Proposition 3.9: Let Assumptions 3.1 and 3.2 hold. Also, let the sequence $\{x_k\}$ be generated by the randomized method (30) with the dynamic stepsize given by Eq. (40).

(a) We have

$$\liminf_{p\to\infty} \sqrt{p}\, E\Big\{f(x_{Mp}) - f^*\Big\} = 0.$$

(b) For a positive scalar ϵ, we have

$$\min_{0\le p\le K} f(x_{Mp}) \le f^* + \epsilon,$$

with probability 1, where K is a random variable such that

$$E\Big\{K\Big\} \le \frac{m^2C^2}{\epsilon^2\underline{\gamma}(2-\overline{\gamma})} \Big(dist(x_0, X^*)\Big)^2. \tag{46}$$

Proof: (a) Assume, to arrive at a contradiction, that for some $\epsilon > 0$

$$\liminf_{p\to\infty} \sqrt{p} E\Big\{f(x_{Mp}) - f^*\Big\} = 2\epsilon.$$

Then for p_0 large enough we have $E\Big\{f(x_{MP}) - f^*\Big\} \ge \frac{\epsilon}{\sqrt{p}}$ for all $p \ge p_0$. Therefore

$$\sum_{p=p_0}^{\infty} E\Big\{(f(x_{Mp}) - f^*)^2\Big\} \ge \epsilon^2 \sum_{p=p_0}^{\infty} \frac{1}{p} = \infty. \tag{47}$$

On the other hand, by using the definition of the stepsize α_k, from Eq. (41) we have for all $p \ge p_0$

$$E\Big\{(dist(x_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} \le \Big(dist(x_{Mp}, X^*)\Big)^2 - \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\Big(f(x_{Mp}) - f^*\Big)^2, \tag{48}$$

so, after taking the total expectation, we see that

$$\sum_{p=0}^{\infty} E\Big\{f(x_{Mp}) - f^*\Big\}^2 < \infty,$$

which contradicts Eq. (47). Hence, we must have

$$\liminf_{p\to\infty} \sqrt{p} E\Big\{f(x_{pM}) - f^*\Big\} = 0.$$

(b) Define a new process $\{\tilde{x}_k\}$ as follows

$$\tilde{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } f(x_{Mp}) \ge f^* + \epsilon, \\ \tilde{y} & \text{otherwise,} \end{cases} \qquad k = Mp, \ldots, M(p+1) - 1,$$

where $\tilde{x}(0) = x(0)$ and $\tilde{y}$ is a fixed vector in X^*. The process $\{\tilde{x}_k\}$ is identical to $\{x_k\}$, except that once x_{Mp} enters the level set

$$L_\epsilon = \{x \in X \mid f(x) < f^* + \epsilon\}$$

the process $\{\tilde{x}_k\}$ terminates at the point $x_{Mp} = \tilde{y}$. Then for the process $\{\tilde{x}_k\}$ it can be seen that for all p [cf. Eq. (48)]

$$\begin{aligned} E\Big\{(dist(\tilde{x}_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} &\leq \Big(dist(\tilde{x}_{Mp}, X^*)\Big)^2 \\ &\quad - \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\Big(f(\tilde{x}_{Mp}) - f^*\Big)^2 \\ &= \Big(dist(\tilde{x}_{Mp}, X^*)\Big)^2 - z_p, \end{aligned} \tag{49}$$

where

$$z_p = \begin{cases} \frac{\gamma_p(2-\gamma_p)\left(f(\tilde{x}_{Mp})-f^*\right)^2}{m^2C^2} & \text{if } \tilde{x}_{Mp} \notin L_\epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

In the case where $\tilde{x}_{Mp} \notin L_\epsilon$, we have

$$\begin{aligned} z_p &\geq \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\Big(f^* + \epsilon - f^*\Big)^2 \\ &\geq \frac{\underline{\gamma}(2-\overline{\gamma})\epsilon^2}{m^2C^2}, \end{aligned} \tag{50}$$

where the last inequality above follows from the fact $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ for all p. By the supermartingale convergence theorem, from Eq. (49) we have

$$\sum_{p=0}^{\infty} z_p < \infty$$

with probability 1, so that $z_p = 0$ for all $p \geq K$, where K is a random variable. Hence $\tilde{x}_{MK} \in L_\epsilon$ with probability 1, implying that in the original process we have

$$\min_{0 \leq p \leq K} f(x_{Mp}) \leq f^* + \epsilon$$

with probability 1. Furthermore, by taking the total expectation in Eq. (49), we obtain for all p

$$\begin{aligned} E\Big\{(dist(\tilde{x}_{M(p+1)}, X^*))^2\Big\} &\leq E\Big\{(dist(\tilde{x}_{Mp}, X^*))^2\Big\} - E\Big\{z_p\Big\} \\ &\leq \Big(dist(x_0, X^*)\Big)^2 - E\left\{\sum_{j=0}^{p} z_j\right\}, \end{aligned}$$

where in the last inequality we use the facts $\tilde{x}_0 = x_0$ and

$$E\Big\{(dist(x_0, X^*))^2\Big\} = \Big(dist(x_0, X^*)\Big)^2.$$

Therefore

$$\Big(dist(x_0, X^*)\Big)^2 \geq E\left\{\sum_{k=0}^{\infty} z_k\right\} = E\left\{\sum_{k=0}^{K-1} z_k\right\} \geq E\Big\{K\Big\}\frac{\underline{\gamma}(2-\overline{\gamma})\epsilon^2}{m^2C^2},$$

where the last inequality above follows from Eq. (50). **Q.E.D.**

Under an additional assumption on f, we can obtain a different estimate of the convergence rate for the method (30) with the dynamic stepsize.

Proposition 3.10: Let Assumptions 3.1 and 3.2 hold. Assume that for some positive scalar μ, with probability 1, we have

$$f(x) - f^* \geq \mu \; dist(x, X^*), \qquad \forall\, x \in X. \tag{51}$$

Then, for a sequence generated by the randomized method (30) with the dynamic stepsize (40) we have

$$E\Big\{dist(x_{Mp}, X^*)\Big\} \leq r^p \; dist(x_0, X^*), \qquad \forall\, p \geq 0,$$

where

$$r = \sqrt{1 - \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}}.$$

Proof: By using the definition of the stepsize α_k, from Eq. (41) we obtain for all $p \geq p_0$

$$E\Big\{(dist(x_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} \leq \Big(dist(x_{Mp}, X^*)\Big)^2 - \gamma_p(2-\gamma_p)\frac{\Big(f(x_{Mp}) - f^*\Big)^2}{m^2C^2}.$$

By taking the total expectation in the above inequality, by using the given property of f [cf. Eq. (51)] and the fact $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ for all p, we have for all p

$$E\Big\{(dist(x_{M(p+1)}, X^*))^2\Big\} \leq \left(1 - \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}\right) E\Big\{(dist(x_{Mp}, X^*))^2\Big\}.$$

The result follows from this relation and the fact

$$E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} = \Big(dist(x_0,X^*)\Big)^2.$$

Q.E.D.

It is difficult to compare the results of Propositions 3.9 and 3.10 with the results of the corresponding Propositions 2.10 and 2.11. Based on these results, if M is much smaller than m, then the convergence rate of the randomized order method is superior. However, for a small M, there is an increased overhead associated with calculating the value of the dynamic stepsize.

3.4 Dynamic Stepsize Rule for Unknown f^*

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

$$\alpha_k = \gamma_p \frac{f_p - f_p^{\rm lev}}{mMC^2}, \qquad 0 < \underline{\gamma} \le \gamma_p \le \overline{\gamma} < 2, \quad k = Mp, \ldots, M(p+1)-1, \tag{52}$$

where $f_p = f(x_{Mp})$. 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

$$f_p^{\rm lev} = \min_{0\le j\le p} f(x_{Mj}) - \delta_p, \tag{53}$$

where δ_p is updated according to

$$\delta_{p+1} = \begin{cases} \delta_p & \text{if } f(x_{M(p+1)}) < f_p^{\rm lev}, \\ \max\{\beta\delta_p, \delta\} & \text{if } f(x_{M(p+1)}) \ge f_p^{\rm lev}, \end{cases} \tag{54}$$

where δ_0, δ and β are fixed positive constants with $\beta < 1$. Note here that we have set to 1 the parameter ρ of Eq. (26). Our results rely on this fact. Since the stepsize is bounded away from zero, the method behaves similar to the one with a constant stepsize (cf. Proposition 3.2). More precisely, we have the following result.

Proposition 3.11: Let Assumption 3.1 hold and let a sequence $\{x_k\}$ be generated by the randomized method (30) with the stepsize (52) and the adjustment procedure (53)–(54).

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

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

(b) If f^* is finite, then with probability 1 we have

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

Proof: See Proposition 3.4 of Nedić and Bertsekas [25]. **Q.E.D.**

We first note that in analogy with Proposition 3.7, we can prove that for all p [cf. Eq. (41) with α_{Mp} as in Eq. (52)]

$$E\Big\{(dist(\tilde{x}_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} \leq \\ \Big(dist(\tilde{x}_{Mp}, X^*)\Big)^2 - \frac{2\gamma_p(f_p - f_p^{\mathrm{lev}})}{m^2C^2}\Big(f_p - f^*\Big) + \frac{\gamma_p^2(f_p - f_p^{\mathrm{lev}})^2}{m^2C^2}, \quad (55)$$

where $f_p = f(\tilde{x}_{Mp}) = f(x_{Mp})$. The next result parallels the one of Proposition 2.13 for the fixed order method.

Proposition 3.12: Let Assumptions 3.1 and 3.2 hold. Then, for a sequence $\{x_k\}$ generated by the randomized method (30) with the stepsize (52) and the adjustment procedure (53)–(54), we have

$$\min_{0\leq p\leq K} f(x_{Mp}) \leq f^* + \delta_0$$

with probability 1, where K is a random variable such that

$$E\Big\{K\Big\} \leq \frac{m^2C^2\Big(dist(x_0, X^*)\Big)^2}{\underline{\gamma}(2-\overline{\gamma})\delta^2}.$$

Proof: Define a new process $\{\tilde{x}_k\}$ as follows

$$\tilde{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } f(x_{Mp}) \geq f^* + \delta_0, \\ \tilde{y} & \text{otherwise,} \end{cases} \qquad k = Mp, \ldots, M(p+1) - 1,$$

where $\tilde{x}(0) = x(0)$ and $\tilde{y}$ is a fixed vector in X^*. The process $\{\tilde{x}_k\}$ is identical to $\{x_k\}$, except that once x_{Mp} enters the level set

$$L_0 = \{x \in X \mid f(x) < f^* + \delta_0\}$$

the process $\{\tilde{x}_k\}$ terminates at the point $x_{Mp} = \tilde{y}$. Then, for the process $\{\tilde{x}_k\}$, from Eq. (55) we have

$$E\Big\{(dist(\tilde{x}_{M(p+1)}, X^*))^2 \mid \mathcal{F}_p\Big\} \le \Big(dist(\tilde{x}_{Mp}, X^*)\Big)^2 - z_p, \tag{56}$$

where

$$z_p = \begin{cases} \frac{2\gamma_p(f_p - f_p^{\text{lev}})}{m^2C^2}\Big(f_p - f^*\Big) - \frac{\gamma_p^2}{m^2C^2}(f_p - f_p^{\text{lev}})^2 & \text{if } \tilde{x}_{Mp} \notin L_0, \\ 0 & \text{otherwise.} \end{cases}$$

In the case where $\tilde{x}_{Mp} \neq \tilde{y}$, we have

$$f(\tilde{x}_{Ms}) \ge f^* + \delta_0, \qquad s = 0, \ldots, p,$$

so that

$$f_p^{\text{lev}} = \min_{0 \le j \le p} f(\tilde{x}_{Mj}) - \delta_p \ge f^* + \delta_0 - \delta_p \ge f^*,$$

where the last inequality above follows from the fact $\delta_0 \ge \delta_k$ for all k. Thus for $\tilde{x}_{Mp} \notin L_0$ (since $\tilde{x}_{Mp} \neq \tilde{y}$) we obtain

$$\begin{aligned} z_p &\ge \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\Big(f_p - f_p^{\text{lev}}\Big)^2 \\ &\ge \frac{\underline{\gamma}(2-\overline{\gamma})\delta^2}{m^2C^2}, \end{aligned} \tag{57}$$

where the last inequality above follows from the facts $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ and $f_p - f_p^{\text{lev}} \ge \delta_p \ge \delta$ for all p. By the supermartingale convergence theorem, from Eq. (56) we have

$$\sum_{p=0}^{\infty} z_p < \infty$$

with probability 1, so that $z_p = 0$ for all $p \ge K$, where K is a random variable. Hence $\tilde{x}_{MK} \in L_0$ with probability 1, implying that in the original process we have

$$\min_{0 \le p \le K} f(x_{Mp}) \le f^* + \delta_0$$

with probability 1. Furthermore, by taking the total expectation in Eq. (56), we obtain for all p

$$\begin{aligned} E\Big\{(dist(\tilde{x}_{M(p+1)}, X^*))^2\Big\} &\le E\Big\{(dist(\tilde{x}_{Mp}, X^*))^2\Big\} - E\Big\{z_p\Big\} \\ &\le \Big(dist(x_0, X^*)\Big)^2 - E\left\{\sum_{j=0}^{p} z_j\right\}, \end{aligned}$$

where in the last inequality we use the facts $\tilde{x}_0 = x_0$ and

$$E\Big\{(dist(x_0, X^*))^2\Big\} = \Big(dist(x_0, X^*)\Big)^2.$$

Therefore

$$\Big(dist(x_0, X^*)\Big)^2 \geq E\left\{\sum_{k=0}^{\infty} z_k\right\} = E\left\{\sum_{k=0}^{K-1} z_k\right\} \geq E\Big\{K\Big\}\frac{\underline{\gamma}(2-\overline{\gamma})\delta^2}{m^2C^2},$$

where the last inequality above follows from Eq. (57). **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, similar to the preceding analysis, we can estimate the expected number of cycles K required for

$$\min_{0\leq p\leq K} f(x_{Mp}) \leq f^* + \delta_0$$

to hold.

References

[1] Allen E., Helgason R., Kennington J., and Shetty B. (1987), "A Generalization of Polyak's Convergence Result for Subgradient Optimization," *Mathematical Programming*, 37, 309–317.

[2] Bazaraa M. S., and Sherali H. D. (1981), "On the Choice of Step Size in Subgradient Optimization," *European Journal of Operational Research*, 7, 380–388.

[3] Bertsekas D. P. (1999), *Nonlinear Programming*, (2nd edition), Athena Scientific, Belmont, Massachusetts.

[4] Bertsekas D. P., and Tsitsiklis J. N. (1996), *Neuro-Dynamic Programming*, Athena Scientific, Belmont, Massachusetts.

[5] Brännlund U. (1993), "On Relaxation Methods for Nonsmooth Convex Optimization," *Doctoral Thesis*, Royal Institute of Technology, Stockholm, Sweden.

[6] Correa R., and Lemaréchal C. (1993), "Convergence of Some Algorithms for Convex Minimization," *Mathematical Programming*, 62, 261–275.

[7] Dem'yanov V. F., and Vasil'ev L. V. (1985), *Nondifferentiable Optimization*, Optimization Software, New York.

[8] Ermoliev Yu. M. (1966), "Methods for Solving Nonlinear Extremal Problems," *Kibernetika*, Kiev, 4, 1–17.

[9] Ermoliev Yu. M. (1969), "On the Stochastic Quasi-gradient Method and Stochastic Quasi-Feyer Sequences," *Kibernetika*, 2, 73–83.

[10] Ermoliev Yu. M. (1976), *Stochastic Programming Methods*, Nauka, Moscow.

[11] Ermoliev Yu. M. (1983), "Stochastic Quasigradient Methods and Their Application to System Optimization," *Stochastics*, 9, 1–36.

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

[13] Goffin J. L. (1980), "The Relaxation Method for Solving Systems of Linear Inequalities," *Mathematics of Operations Research*, 5(3), 388–414.

[14] Goffin J., and Kiwiel K. (1999), "Convergence of a Simple Subgradient Level Method," *Mathematical Programming*, 85, 207–211.

[15] Hiriart-Urruty J.-B., and Lemaréchal C. (1993), *Convex Analysis and Minimization Algorithms*, Vols. I and II, Springer-Verlag, Berlin and New York.

[16] Kaskavelis C. A., and Caramanis M. C. (1998), "Efficient Lagrangian Relaxation Algorithms for Industry Size Job-Shop Scheduling Problems," *IIE Transactions on Scheduling and Logistics*, 30, 1085-1097.

[17] Kim S., Ahn H., and Cho S.-C. (1991), "Variable Target Value Subgradient Method," *Mathematical Programming*, 49, 359–369.

[18] Kim S., and Um B. (1993), "An Improved Subgradient Method for Constrained Nondifferentiable Optimization," *Operations Research Letters*, 14, 61–64.

[19] Kiwiel K. C. (1996), "The Efficiency of Subgradient Projection Methods for Convex Optimization, Part I: General Level Methods," *SIAM Journal on Control and Optimization*, 34(2), 660–676.

[20] Kiwiel K. C. (1996), "The Efficiency of Subgradient Projection Methods for Convex Optimization, Part II: Implementations and Extensions," *SIAM Journal on Control and Optimization*, 34(2), 677–697.

[21] Kiwiel K. C., Larsson T., and Lindberg P. O. (1998), "The Efficiency of Ballstep Subgradient Level Methods for Convex Optimization," *Working Paper LiTH-MAT-R-1998-22*, Dept. of Mathematics, Linköpings Universitet, Sweden.

[22] Kulikov A. N., and Fazylov V. R. (1990), "Convex Optimization with Prescribed Accuracy," *USSR Computational Mathematics and Mathematical Physics*, 30(3), 16–22.

[23] Minoux M. (1986), *Mathematical Programming: Theory and Algorithms*, J. Wiley, New York.

[24] Nedić A., and Bertsekas D. P. (1999), "Incremental Subgradient Methods for Nondifferentiable Optimization," *Lab. for Info. and Decision Systems Report LIDS-P-2460*, Massachusetts Institute of Technology, Cambridge, MA.

[25] Nedić A., and Bertsekas D. P. (2000), "Incremental Subgradient Methods for Nondifferentiable Optimization," (to appear in *SIAM Journal on Optimization*).

[26] Polyak B. T. (1967), "A General Method of Solving Extremum Problems," *Doklady Akad. Nauk SSSR*, 174(1), 33–36.

[27] Polyak B. T. (1969), "Minimization of Unsmooth Functionals," *Zhurnal Vychisl. Mat. i Mat. Fiz.*, 9(3), 509–521.

[28] Polyak B. T. (1987), *Introduction to Optimization*, Optimization Software Inc., New York.

[29] Rockafellar R. T. (1970), *Convex Analysis*, Princeton Univ. Press, Princeton, New Jersey.

[30] Shor N. Z. (1985), *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin.

[31] Solodov M. V., and Zavriev S. K. (1998), "Error Stability Properties of Generalized Gradient-Type Algorithms," *Journal of Optimization Theory and Applications*, 98(3), 663–680.

[32] Zhao X., Luh P. B., and Wang J. (1999), "Surrogate Gradient Algorithm for Lagrangian Relaxation," *Journal of Optimization Theory and Applications*, 100(3), 699–712.

Acknowledgment

This work is supported by the National Science Foundation grant ACI-9873339.