

# The effect of deterministic noise in subgradient methods

Angelia Nedić · Dimitri P. Bertsekas

Received: 25 September 2007 / Accepted: 4 December 2008  
© Springer-Verlag 2008

**Abstract** In this paper, we study the influence of noise on subgradient methods for convex constrained optimization. The noise may be due to various sources, and is manifested in inexact computation of the subgradients and function values. Assuming that the noise is deterministic and bounded, we discuss the convergence properties for two cases: the case where the constraint set is compact, and the case where this set need not be compact but the objective function has a sharp set of minima (for example the function is polyhedral). In both cases, using several different stepsize rules, we prove convergence to the optimal value within some tolerance that is given explicitly in terms of the errors. In the first case, the tolerance is nonzero, but in the second case, the optimal value can be obtained exactly, provided the size of the error in the subgradient computation is below some threshold. We then extend these results to objective functions that are the sum of a large number of convex functions, in which case an incremental subgradient method can be used.

**Mathematics Subject Classification (2000)** 90C25

---

Research supported by NSF under Grant ACI-9873339.

---

A. Nedić (✉)  
Department of Industrial and Enterprise Systems Engineering,  
UIUC, Urbana, IL 61801, USA  
e-mail: angelia@uiuc.edu

D. P. Bertsekas  
Department of Electrical Engineering and Computer Science,  
M.I.T., Cambridge, MA 02139, USA

## 1 Introduction

We focus on the problem

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x \in X, \end{aligned} \tag{1.1}$$

where  $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$  is a convex function, and  $X$  is a nonempty, closed, and convex set in  $\mathfrak{R}^n$ . We are primarily interested in the case where  $f$  is nondifferentiable. Throughout the paper, we denote

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

where  $\|\cdot\|$  is the standard Euclidean norm. In our notation, all vectors are assumed to be column vectors and a prime denotes transposition.

We focus on an approximate  $\epsilon$ -subgradient method where the  $\epsilon$ -subgradients are computed inexactly. In particular, the method is given by

$$x_{k+1} = \mathcal{P}_X[x_k - \alpha_k \tilde{g}_k], \tag{1.2}$$

where  $\mathcal{P}_X$  denotes the projection on the set  $X$ . The vector  $x_0$  is an initial iterate from the set  $X$  (i.e.,  $x_0 \in X$ ) and the scalar  $\alpha_k$  is a positive stepsize. The vector  $\tilde{g}_k$  is an approximate subgradient of the following form

$$\tilde{g}_k = g_k + r_k, \tag{1.3}$$

where  $r_k$  is a noise vector and  $g_k$  is an  $\epsilon_k$ -subgradient of  $f$  at  $x_k$  for some  $\epsilon_k \geq 0$ , i.e.,  $g_k$  satisfies

$$f(y) \geq f(x_k) + g_k'(y - x_k) - \epsilon_k, \quad \forall y \in \mathfrak{R}^n. \tag{1.4}$$

We consider several stepsize rules, including a rule using function values  $f(x_k)$ . For this rule, we assume that the function values  $f(x_k)$  are evaluated approximately, and are replaced by  $\tilde{f}(x_k)$ , where

$$\tilde{f}(x_k) = f(x_k) + \xi_k, \quad \forall k \geq 0, \tag{1.5}$$

and  $\xi_k$  is some scalar error.

We quantify the joint effect of the noise level<sup>1</sup> ( $\sup_k \|r_k\|$ ), the approximate-subgradient error level ( $\limsup_k \epsilon_k$ ), and the function value error ( $\sup_k |\xi_k|$ ). In particular, we study the convergence properties of the method (1.2) using the following stepsize rules:

<sup>1</sup> In our subsequent development, the  $\sup_k \|r_k\|$  and  $\sup_k \|r_k\|$  and  $\sup_k |\xi_k|$  can be replaced with  $\limsup_{k \rightarrow \infty} \|r_k\|$  and  $\limsup_{k \rightarrow \infty} |\xi_k|$  with minor technical adjustments in the proofs of the results of Sect. 2, and some additional care in the proofs of the results of Sect. 3 and 4.

- (a) *Constant stepsize rule.* The stepsize  $\alpha_k$  is fixed to a positive scalar  $\alpha$ .
- (b) *Diminishing stepsize rule.* The stepsize  $\alpha_k > 0$  satisfies

$$\lim_{k \rightarrow \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

- (c) *Dynamic stepsize rule.* The stepsize is given by

$$\alpha_k = \gamma_k \frac{\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}}{\|\tilde{g}_k\|^2}, \quad 0 < \underline{\gamma} \leq \gamma_k \leq 2, \quad \forall k \geq 0, \quad (1.6)$$

where  $\tilde{f}(x_k)$  is an error-corrupted function value as in Eq. (1.5), while the scalars  $\tilde{f}_k^{\text{lev}}$  are the target levels approximating the optimal value  $f^*$ . For this stepsize rule, we consider two procedures for adjusting the target levels  $\tilde{f}_k^{\text{lev}}$  (cf. Sect. 2).

The issue of noise in the context of subgradient optimization was first studied by Ermoliev in [7] (see also Ermoliev [8, 9], and [10], and Nurminskii [21]), where a random noise was considered. When the noise is deterministic, the stochastic subgradient method analyzed by Ermoliev is similar to the special case of method (1.2) with the diminishing stepsize  $\alpha_k$ , and the diminishing noise  $r_k$  and zero  $\epsilon_k$ -errors (i.e.,  $\epsilon_k \equiv 0$ ). In this case, the convergence of the method is not affected by the presence of noise as long as the stepsize  $\alpha_k$  and the noise magnitude  $\|r_k\|$  are coordinated. The presence of (deterministic and stochastic) noise in subgradient methods was also addressed by Polyak in [23] and [24], where the focus is on conditions under which the convergence to the optimal value  $f^*$  is preserved. In Polyak’s work, the convergence of method (1.2) with  $\epsilon_k \equiv 0$  is studied for diminishing stepsize and for a stepsize rule due to Shor that has the form  $\alpha_k = \alpha_0 q^k$ , where  $\alpha_0 > 0$  and  $0 < q < 1$  (see Theorem 4 of [23], or Theorem 1 in Sect. 5 of Chapter 5 in [24]). An interesting result is shown for Shor’s stepsize, under the assumption that the function  $f$  has a unique sharp minimum  $x^*$ . The result shows exact convergence of the method even when the subgradient noise  $r_k$  is nonvanishing. Specifically, when the noise magnitude is “small enough” (with respect to the “sharpness” of  $f$ ) and the initial stepsize value  $\alpha_0$  is proportional to the distance  $\|x_0 - x^*\|$ , the iterates  $x_k$  of the method converge linearly to the optimal vector  $x^*$ . There is also related work of Solodov and Zavriev [26], where a subgradient method and its various modifications were considered in the presence of bounded noise. This work addresses a more general class of objective functions (including nonconvex), but is restricted to a compact constraint set  $X$  and focused on algorithms using only diminishing stepsize. The effects of noise for a proximal bundle method has been recently studied by Kiwiel [16], who has considered evaluating the objective function and its subgradients with a fixed (but possibly unknown) error  $\epsilon$ . The aforementioned works (except for [26] and [16]) are primarily focused on studying the diminishing errors and the necessary conditions on these errors guaranteeing convergence of the subgradient methods in the presence of noise.

By contrast, in this paper we are primarily concerned with cases where the noise and subgradient approximation errors are persistent (nondiminishing). Our main objective is to obtain error bounds on the difference between the attained function value and

the optimal,  $\inf_{k \geq 0} f(x_k) - f^*$ , under a variety of noise conditions and stepsize rules. One contribution of our work is the establishment of error bounds under a richer set of error conditions, including simultaneous errors in the  $\epsilon$ -subgradient and function value computations. Another contribution is in quantifying the effects of errors and noise in conjunction with the use of constant and dynamic stepsize rules. While these stepsize rules have been used in  $\epsilon$ -subgradient methods where  $r_k \equiv 0$ ,  $\xi_k \equiv 0$ , and  $\epsilon_k > 0$ , they have not been considered for cases where  $r_k \neq 0$  and/or  $\xi_k \neq 0$ . A third contribution is the study of the effects of noise in the presence of a sharp minimum and for the stepsize rules that we consider. We finally note that the effects of errors in the context of incremental subgradient methods have been studied earlier only for  $\epsilon$ -subgradients by Solodov and Zavriev [26], and Kiwiel [15] (where  $r_k \equiv 0$ ,  $\xi_k \equiv 0$ , and  $\epsilon_k > 0$ ).

The motivation to study the methods with subgradient errors of the form (1.3)–(1.5), where  $\|r_k\| > 0$ ,  $\epsilon_k > 0$ , and  $|\xi_k| > 0$  comes from several contexts. A common situation arises in optimization of dual functions, where the dual function value is computed with an inexact Lagrangian minimization (within  $\epsilon$ ) that yields an  $\epsilon$ -accurate dual function value and an  $\epsilon$ -subgradient (see e.g., Bertsekas [2], Sect. 6.3.2). Another interesting context comes from more recent network applications where the transmitted data is quantized, as discussed in the following subsection.

### 1.1 Motivating example

Consider distributed optimization in networks consisting of  $m$  nodes and a fusion center, as discussed for example in Nedić, Bertsekas, and Borkar [17] (see also Nedić [20]). Each node  $i$  has an objective function  $f_i$  known only at that node, while the global system objective is to minimize  $f(x) = \sum_{i=1}^m f_i(x)$  over a constraint set  $X$ . The fusion center is responsible for updating  $x_k$  and broadcasting this estimate to the nodes in the network. In return, upon receiving  $x_k$ , each node  $i$  computes a subgradient of its objective  $f_i$  at  $x_k$  and sends the subgradient information to the fusion center. However, in many applications, the communications links between the fusion center and the nodes can handle only quantized data, (see for example, Rabbat and Nowak [25], Kashyap, Basar, and Srikant [13], Tuncer, Coates, and Rabbat [27]). When the quantization level of the links is  $Q$  (a positive integer), each coordinate of the estimate  $x_k$  is quantized with respect to the level  $Q$ . Thus, the nodes receive the quantized estimate  $x_k^Q$  instead of the true estimate  $x_k$ . Each node  $i$  computes a subgradient  $g_{i,k}$  of  $f_i$  at  $x_k^Q$  and sends it to the fusion center. Again, due to the quantization of the transmitted data, the fusion center receives the quantized subgradients  $g_{1k}^Q, \dots, g_{mk}^Q$  and performs the iterate update using these subgradients. In this case, the vector  $\tilde{g}_k$  in Eq. (1.2) is given by  $\tilde{g}_k = g_k^Q = \sum_{i=1}^m g_{ik}^Q$ , and it can be written as follows

$$\tilde{g} = g_k + r_k, \quad \text{with} \quad g_k = \sum_{i=1}^m g_{ik}, \quad r_k = \sum_{i=1}^m (g_{ik}^Q - g_{ik}),$$

where  $g_{ik}$  is a subgradient of  $f_i$  at  $x_k^Q$ . Hence,  $g_k \in \partial f(x_k^Q)$  where  $f = \sum_{i=1}^m f_i$ . By using  $\epsilon$ -subgradient inequality (1.4), it can be seen that  $g_k$  is an  $\epsilon_k$ -subgradient of  $f$

at  $x_k$  with  $\epsilon_k = |f(x_k) - f(x_k^Q) - g'_k(x_k - x_k^Q)|$ . The noise  $r_k$  is deterministic and it is due to quantization of the subgradients  $g_{1k}, \dots, g_{mk}$  of the functions  $f_1, \dots, f_m$  at  $x_k^Q$ . The error  $\epsilon_k$  and noise norm  $\|r_k\|$  can be related<sup>2</sup> to the quantization level  $Q$ , given the quantization mechanism. For example, when the quantization is performed by rounding to the closest integer multiple of  $1/Q$ , the resulting errors  $r_k$  have norm bounded by a constant  $m\sqrt{n}/(2Q)$ , where  $n$  is the size of the vector  $x$  and  $m$  is the number of nodes. The persistent errors in function evaluation of the form (1.5) arise when the nodes communicate their objective function values to the fusion center.

### 1.2 Paper organization

This paper is organized as follows: In Sect. 2, we give the convergence properties of the method for a compact constraint set  $X$ .<sup>3</sup> In Sect. 3, we discuss the convergence properties of the method for the case when the objective function  $f$  has a set of sharp minima (also known as weak sharp minima, see e.g., Burke and Ferris [5]). As a special case, our results show that with  $\epsilon_k \equiv 0$  and the diminishing stepsize rule, the method converges to the optimal value  $f^*$  even if the noise is nonvanishing but is instead small enough (relative to the “sharpness” of the set of minima).<sup>4</sup> In Sect. 4, we consider an objective function  $f$  that is the sum of a large number of convex functions, in which case an incremental subgradient method can also be used. We give analogs of the results of Sects. 2 and 3 for incremental subgradient methods.

## 2 Convergence properties for a compact $X$

In this section we discuss the convergence properties of the method for the case when the constraint set  $X$  is compact. In the following lemma, we give a basic relation that holds for the iterates  $x_k$  obtained by using any of the stepsize rules described in Sect. 1.

**Lemma 2.1** *Let  $X^*$  be nonempty. Then, for a sequence  $\{x_k\}$  generated by the method with any of the stepsize rules (a)–(c), we have for all  $k$*

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k(f(x_k) - f^*) + 2\alpha_k\epsilon_k \\ &\quad + 2\alpha_k\|r_k\|\text{dist}(x_k, X^*) + \alpha_k^2\|\tilde{g}_k\|^2. \end{aligned}$$

<sup>2</sup> Note that the total error  $\epsilon_k + \|r_k\|$  resulting from using the quantized vector  $g_k^Q$  cannot be bundled into a “larger”  $\epsilon$ -subgradient type error. This is because  $\epsilon$ -subgradient errors at a given  $x_k$  arise from approximating  $x_k$  with some  $x$  in a neighborhood of  $x_k$ . In contrast, the vector  $g_k^Q$  arises from two approximations: approximating  $x_k$  (by using a “nearby” point  $x_k^Q$ ) and approximating a subgradient  $g_k$  of  $f$  at  $x_k^Q$  (by using a “nearby” direction  $g_k^Q$ ).

<sup>3</sup> The results of Sect. 2 actually hold under the weaker assumption that the optimal set  $X^*$  is nonempty, and the sequences  $\{g_k\}$  and  $\{\text{dist}(x_k, X^*)\}$  are bounded. The principal case where this is guaranteed without assuming compactness of  $X$  is when  $f$  is polyhedral, and  $X^*$  is nonempty and bounded. The case of polyhedral  $f$ , however, is treated separately in Sect. 3, so for simplicity, in Sect. 2 we assume that  $X$  is compact.

<sup>4</sup> Our result for diminishing stepsize has been established by Solodov and Zavriev in [26], Lemma 4.3, under the additional assumption that the constraint set is compact.

*Proof* Using the definition of  $x_{k+1}$  in Eq. (1.2) and the nonexpansion property of the projection, we obtain for all  $y \in X$ ,

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y - \alpha_k \tilde{g}_k\|^2 \\ &= \|x_k - y\|^2 - 2\alpha_k \tilde{g}'_k(x_k - y) + \alpha_k^2 \|\tilde{g}_k\|^2 \\ &\leq \|x_k - y\|^2 - 2\alpha_k g'_k(x_k - y) + 2\alpha_k \|\tilde{g}_k - g_k\| \cdot \|x_k - y\| + \alpha_k^2 \|\tilde{g}_k\|^2 \\ &\leq \|x_k - y\|^2 - 2\alpha_k (f(x_k) - f(y)) + 2\alpha_k \epsilon_k + 2\alpha_k \|r_k\| \\ &\quad \cdot \|x_k - y\| + \alpha_k^2 \|\tilde{g}_k\|^2, \end{aligned}$$

where in the last inequality we use the  $\epsilon_k$ -subgradient property (1.4), and the fact  $\tilde{g}_k - g_k = r_k$  [cf. Eq. (1.3)]. The desired relation follows from the preceding inequality by letting  $y = \mathcal{P}_{X^*}[x_k]$ , and by using the relations

$$\|x_k - \mathcal{P}_{X^*}[x_k]\| = \text{dist}(x_k, X^*), \quad \text{dist}(x_{k+1}, X^*) \leq \|x_{k+1} - \mathcal{P}_{X^*}[x_k]\|.$$

□

Throughout this section, we use the following assumptions.

**Assumption 2.1** The constraint set  $X$  is compact.

**Assumption 2.2** The noise  $r_k$  and the errors  $\epsilon_k$  are bounded, i.e., for some scalars  $R \geq 0$  and  $\epsilon \geq 0$  there holds

$$\|r_k\| \leq R, \quad \forall k \geq 0, \quad \text{and} \quad \limsup_{k \rightarrow \infty} \epsilon_k = \epsilon.$$

When the set  $X$  is compact (cf. Assumption 2.1), the optimal set  $X^*$  is nonempty, and the sequences  $\{g_k\}$  and  $\{\text{dist}(x_k, X^*)\}$  are bounded. Hence, for some positive scalars  $C$  and  $d$ , we have

$$\|g_k\| \leq C, \quad \text{dist}(x_k, X^*) \leq d, \quad \forall k \geq 0. \quad (2.1)$$

Furthermore, under bounded noise (cf. Assumption 2.2), from the relation  $\tilde{g}_k = g_k + r_k$  [cf. Eq. (1.3)] it follows that the directions  $\tilde{g}_k$  are uniformly bounded

$$\|\tilde{g}_k\| \leq C + R, \quad \forall k \geq 0. \quad (2.2)$$

Note that under the compactness assumption, a simple bound on the distance between the iterates  $x_k$  and the optimal set  $X^*$  can be obtained by letting  $d$  in Eq. (2.1) be equal to the diameter of the set  $X$  (i.e.,  $d = \max_{x, y \in X} \|x - y\|$ ). More complex and tighter bounds may be obtained using the existing error bound results discussed in Pang [22] and the literature cited therein.

We now give the convergence properties for each of the stepsize rules described in Sect. 1. We start with the constant stepsize rule for which we have the following result.

**Proposition 2.1** *Let Assumptions 2.1 and 2.2 hold. Then, for a sequence  $\{x_k\}$  generated by the method with the constant stepsize rule, we have*

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \epsilon + Rd + \frac{\alpha}{2}(C + R)^2.$$

*Proof* In order to arrive at a contradiction, assume that

$$\liminf_{k \rightarrow \infty} f(x_k) > f^* + \epsilon + Rd + \frac{\alpha}{2}(C + R)^2,$$

so that for some nonnegative integer  $k_0$  and a positive scalar  $\nu$  we have

$$f(x_k) \geq f^* + \epsilon_k + Rd + \frac{\alpha}{2}(C + R)^2 + \nu, \quad \forall k \geq k_0.$$

Next, by using Lemma 2.1 with  $\alpha_k = \alpha$ , and the bounds on  $\|r_k\|$ ,  $\text{dist}(x_k, X^*)$ , and  $\|\tilde{g}_k\|$  [cf. Eqs. (2.1) and (2.2)], we obtain for all  $k$

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha \left( f(x_k) - f^* - \epsilon_k - Rd - \frac{\alpha}{2}(C + R)^2 \right).$$

By combining the preceding two relations, we have for all  $k \geq k_0$

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha\nu \\ &\leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 - 2(k + 1 - k_0)\alpha\nu, \end{aligned}$$

which yields a contradiction for sufficiently large  $k$ . □

As suggested by Proposition 2.1, we expect that the error term involving the stepsize  $\alpha$  diminishes to zero as  $\alpha \rightarrow 0$ . Indeed this is so, as shown in the following proposition.

**Proposition 2.2** *Let Assumptions 2.1 and 2.2 hold. Then, for a sequence  $\{x_k\}$  generated by the method with the diminishing stepsize rule, we have*

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \epsilon + Rd.$$

*Proof* The proof uses the fact  $\sum_{k=0}^{\infty} \alpha_k = \infty$  and a line of analysis similar to that of Proposition 2.1. □

We next consider the dynamic stepsize rule of Eq. (1.6). For this rule, we assume that the function value error  $\xi_k = \tilde{f}(x_k) - f(x_k)$  is uniformly bounded, as given in the following.

**Assumption 2.3** The function errors  $\xi_k$  are bounded by a scalar  $\xi \geq 0$ , i.e.,  $|\xi_k| \leq \xi$  for all  $k$ .

We study two adjustment procedures for generating the target level  $\tilde{f}_k^{\text{lev}}$ , one resulting in a nonvanishing stepsize  $\alpha_k$  and the other one resulting in a vanishing  $\alpha_k$ . We start with the one with nonvanishing  $\alpha_k$ . In this procedure, adapted from a stepsize rule considered in [19], the target level  $\tilde{f}_k^{\text{lev}}$  is given by

$$\tilde{f}_k^{\text{lev}} = \min_{0 \leq j \leq k} \tilde{f}(x_j) - \delta_k, \tag{2.3}$$

where  $\delta_k$  is a positive scalar that is updated as follows:

$$\delta_{k+1} = \begin{cases} \bar{\beta}\delta_k & \text{if } \tilde{f}(x_{k+1}) \leq \tilde{f}_k^{\text{lev}}, \\ \max\{\underline{\beta}\delta_k, \delta\} & \text{if } \tilde{f}(x_{k+1}) > \tilde{f}_k^{\text{lev}}, \end{cases} \tag{2.4}$$

where  $\bar{\beta}$ ,  $\underline{\beta}$ ,  $\delta_0$ , and  $\delta$  are fixed positive scalars, with  $\bar{\beta} \geq 1$ ,  $\underline{\beta} < 1$ , and  $\delta_0 \geq \delta$ . Note that in this procedure, we always have  $\delta_k \geq \delta$ . Furthermore, if we set  $\delta_0 = \delta$  and  $\bar{\beta} = 1$ , then  $\delta_k = \delta$  for all  $k$ . Therefore the procedure includes, as a special case, a procedure where  $\delta_k$  is fixed to a positive constant.

Since the procedure (2.3)–(2.4) gives a nonvanishing stepsize  $\alpha_k$ , the convergence property of the method is similar to that of a constant stepsize, as shown in the following.

**Proposition 2.3** *Let Assumptions 2.1–2.3 hold. Then, for a sequence  $\{x_k\}$  generated by the method and the dynamic stepsize rule using the adjustment procedure (2.3)–(2.4), we have*

$$\inf_{k \geq 0} f(x_k) \leq f^* + 2\xi + \epsilon + Rd + \delta.$$

*Proof* To arrive at a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) > f^* + 2\xi + \epsilon + Rd + \delta.$$

We have  $\tilde{f}(x_k) = f(x_k) + \xi_k \geq f(x_k) - |\xi_k|$ , implying (by Assumption 2.3) that  $\tilde{f}(x_k) \geq f(x_k) - \xi$ . Therefore, it follows that

$$\inf_{k \geq 0} \tilde{f}(x_k) > f^* + \xi + \epsilon + Rd + \delta. \tag{2.5}$$

According to the procedure in Eqs. (2.3)–(2.4), we have  $\delta_k \geq \delta$  for all  $k$ . Therefore, each time the target level is attained [i.e.,  $\tilde{f}(x_k) \leq \tilde{f}_{k-1}^{\text{lev}}$ ], the best current function value  $\min_{0 \leq j \leq k} \tilde{f}(x_j)$  is decreased by at least  $\delta$ . In view of this and the relation in Eq. (2.5), the target level can be attained only a finite number of times. From Eq. (2.4) it follows that after finitely many iterations,  $\delta_k$  is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is a nonnegative integer  $k_0$  such that



$$\delta_k = \delta, \quad \forall k \geq k_0.$$

By using the relations  $\tilde{f}_k^{\text{lev}} = \min_{0 \leq j \leq k} \tilde{f}(x_j) - \delta$  for  $k \geq k_0$  and  $\xi \geq \xi_k$  for all  $k$ , and by choosing a larger  $k_0$  if necessary, from Eq. (2.5) it can be seen that for some positive scalar  $\nu$  we have

$$\tilde{f}_k^{\text{lev}} - f^* \geq \xi_k + \epsilon_k + Rd + \nu, \quad \forall k \geq k_0. \tag{2.6}$$

Next, by using Lemma 2.1, and the bounds on  $\|r_k\|$  and  $\text{dist}(x_k, X^*)$ , we obtain for all  $k$

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (f(x_k) - f^*) + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k Rd + \alpha_k^2 \|\tilde{g}_k\|^2. \end{aligned}$$

By writing  $f(x_k) - f^* = \tilde{f}(x_k) - f^* - \xi_k$  and then, by adding and subtracting  $\tilde{f}_k^{\text{lev}}$ , we obtain

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}) \\ &\quad - 2\alpha_k (\tilde{f}_k^{\text{lev}} - f^* - \xi_k) + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k Rd + \alpha_k^2 \|\tilde{g}_k\|^2 \\ &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (\tilde{f}_k^{\text{lev}} - f^* - \xi_k - \epsilon_k - Rd), \end{aligned} \tag{2.7}$$

where the last inequality in the preceding relation follows from the definition of  $\alpha_k$  and the following relation

$$-2\alpha_k (\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}) + \alpha_k^2 \|\tilde{g}_k\|^2 = -\gamma_k (2 - \gamma_k) \frac{(\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}})^2}{\|\tilde{g}_k\|^2} \leq 0, \quad \forall k \geq 0.$$

By using inequality (2.6) in relation (2.7), we have

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \nu \leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 \\ &\quad - 2\nu \sum_{j=k_0}^k \alpha_j, \quad \forall k \geq k_0. \end{aligned} \tag{2.8}$$

Since  $\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}} \geq \delta$  and  $\|\tilde{g}_k\| \leq C + R$  for all  $k$  [cf. Eq. (2.2)], from the definition of  $\alpha_k$  it follows that

$$\alpha_k \geq \underline{\gamma} \frac{\delta}{(C + R)^2}, \quad \forall k \geq k_0,$$

which when substituted in Eq. (2.8) yields a contradiction for sufficiently large  $k$ .  $\square$

In the following algorithm, we describe a path-based procedure for adjusting the target levels  $\tilde{f}_k^{\text{lev}}$ . This procedure is based on the algorithm of Brännlund [4], which was further developed by Goffin and Kiwiel [12].

*The path-based procedure*

- Step 0** (*Initialization*) Select  $x_0 \in X$ ,  $\delta_0 > 0$ , and  $B > 0$ . Set  $\sigma_0 = 0$ ,  $\tilde{f}_{-1}^{\text{rec}} = \infty$ . Set  $k = 0$ ,  $l = 0$ , and  $k(l) = 0$  [ $k(l)$  will denote the iteration number when the  $l$ -th update of  $\tilde{f}_k^{\text{lev}}$  occurs].
- Step 1** (*Function evaluation*) Calculate  $\tilde{f}(x_k)$  and  $\tilde{g}_k$ , where  $\tilde{g}_k$  is given by Eq. (1.3). If  $\tilde{f}(x_k) < \tilde{f}_{k-1}^{\text{rec}}$ , then set  $\tilde{f}_k^{\text{rec}} = \tilde{f}(x_k)$ . Otherwise set  $\tilde{f}_k^{\text{rec}} = \tilde{f}_{k-1}^{\text{rec}}$  [so that  $\tilde{f}_k^{\text{rec}}$  keeps the record of the smallest value attained by the iterates that are generated so far, i.e.,  $\tilde{f}_k^{\text{rec}} = \min_{0 \leq j \leq k} \tilde{f}(x_j)$ ].
- Step 2** (*Sufficient descent*) If  $\tilde{f}(x_k) \leq \tilde{f}_{k(l)}^{\text{rec}} - \frac{\delta_l}{2}$ , 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  $\tilde{f}_k^{\text{lev}} = \tilde{f}_{k(l)}^{\text{rec}} - \delta_l$ . Select  $\gamma_k \in [\underline{\gamma}, 2]$  and calculate  $x_{k+1}$  via Eq. (1.2) with the stepsize (1.6).
- Step 5** (*Path length update*) Set  $\sigma_{k+1} = \sigma_k + \alpha_k \|\tilde{g}_k\|$ . Increase  $k$  by 1 and go to Step 1.

The algorithm uses the same target level  $\tilde{f}_k^{\text{lev}} = \tilde{f}_{k(l)}^{\text{rec}} - \delta_l$  for  $k = k(l), k(l) + 1, \dots, k(l+1) - 1$ . 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  $\sigma_k$  is an upper bound on the length of the path traveled by iterates  $x_{k(l)}, \dots, x_k$  for  $k < k(l+1)$ . If the target level  $\tilde{f}_k^{\text{lev}}$  is too low (i.e., sufficient descent cannot occur), then due to oscillations of  $x_k$  the parameter  $\sigma_k$  eventually exceeds the prescribed upper bound  $B$  on the path length and the parameter  $\delta_l$  is decreased.

We have the following result for the path-based procedure.

**Proposition 2.4** *Let Assumptions 2.1–2.3 hold. Then, for a sequence generated by the method and the path-based procedure, we have*

$$\inf_{k \geq 0} f(x_k) \leq f^* + 2\xi + \epsilon + Rd.$$

*Proof* At first, we show that  $l \rightarrow \infty$ . Suppose that  $l$  takes only a finite number of values, say  $l = 0, 1, \dots, \bar{l}$ , then

$$\sigma_k + \alpha_k \|\tilde{g}_k\| = \sigma_{k+1} \leq B, \quad \forall k \geq k(\bar{l}),$$

so that  $\lim_{k \rightarrow \infty} \alpha_k \|\tilde{g}_k\| = 0$ . But this is impossible, since

$$\alpha_k \|\tilde{g}_k\| = \gamma_k \frac{\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}}{\|\tilde{g}_k\|} \geq \underline{\gamma} \frac{\delta_{\bar{l}}}{2(C + R)}, \quad \forall k \geq k(\bar{l}).$$

Hence  $l \rightarrow \infty$ .

Now, to arrive at a contradiction, assume that  $\inf_{k \geq 0} f(x_k) > f^* + 2\xi + \epsilon + Rd$ . Then, in view of  $\tilde{f}(x_k) = f(x_k) + \xi_k$  and  $|\xi_k| \leq \xi$  for all  $k$ , it follows that

$$\inf_{k \geq 0} \tilde{f}(x_k) > f^* + \xi + \epsilon + Rd. \tag{2.9}$$

If  $\delta_l$  is decreased at Step 3 only a finite number of times, then there must be an infinite number of sufficient descents, so that for some nonnegative integer  $\bar{l}$  we have  $\delta_l = \delta_{\bar{l}} > 0$  for all  $l \geq \bar{l}$ . Each time a sufficient descent is detected, the current best function value  $\min_{0 \leq j \leq k} \tilde{f}(x_j)$  is decreased by at least  $\delta_{\bar{l}}/2$ , so in view of Eq. (2.9) there can be only a finite number of sufficient descents, which is a contradiction. Therefore  $\delta_l$  must be decreased at Step 3 infinitely often, i.e.,  $\lim_{l \rightarrow \infty} \delta_l = 0$ , so that for a sufficiently large positive integer  $\bar{l}$  and a positive scalar  $\nu$  we have [cf. Eq. (2.9)]

$$\inf_{j \geq 0} \tilde{f}(x_j) - \delta_l - f^* \geq \xi + \epsilon_k + Rd + \nu, \quad \forall k \geq k(l), \quad \forall l \geq \bar{l}.$$

Consequently [since  $\tilde{f}_k^{\text{lev}} = \min_{0 \leq j \leq k(l)} \tilde{f}(x_j) - \delta_l$  and  $\xi \geq \xi_k$ ]

$$\tilde{f}_k^{\text{lev}} - f^* \geq \xi_k + \epsilon_k + Rd + \nu, \quad \forall k \geq k(\bar{l}).$$

Similar to the proof of Proposition 2.3, it can be seen that for all  $k$  we have [cf. Eq. (2.7)]

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (\tilde{f}_k^{\text{lev}} - f^* - \xi_k - \epsilon_k - Rd),$$

from which, by using the preceding relation, we obtain for all  $k \geq k(\bar{l})$

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \nu \leq \dots \leq (\text{dist}(x_{k(\bar{l})}, X^*))^2 - 2\nu \sum_{j=k(\bar{l})}^k \alpha_j.$$

Hence  $\sum_{k=0}^{\infty} \alpha_k$  is finite.

Let  $L$  be given by

$$L = \left\{ l \in \{1, 2, \dots\} \mid \delta_l = \frac{\delta_{l-1}}{2} \right\}.$$

Then from Steps 3 and 5 we have

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} \|\tilde{g}_{k-1}\| = \sum_{j=k(l)}^{k-1} \alpha_j \|\tilde{g}_j\|, \tag{2.10}$$

so that, whenever  $\sum_{j=k(l)}^{k-1} \alpha_j \|\tilde{g}_j\| > B$  at Step 3, we have  $k(l+1) = k$  and  $l+1 \in L$ . Therefore

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j \|\tilde{g}_j\| > B, \quad \forall l \in L,$$

which combined with the fact  $\|\tilde{g}_k\| \leq C + R$  for all  $k$  [cf. Eq. (2.2)] implies that for all  $l \in L$

$$(C + R) \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > B.$$

By summing, since the cardinality of  $L$  is infinite, we obtain

$$\sum_{k=0}^{\infty} \alpha_k \geq \sum_{l \in L} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \sum_{l \in L} \frac{B}{C + R} = \infty,$$

which contradicts the finiteness of  $\sum_{k=0}^{\infty} \alpha_k$ . □

Note that the results of Propositions 2.3 and 2.4 have twice the error  $\xi$  coming from the erroneous function evaluations. This can be attributed to the fact that the error in the difference  $\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}$  can reach  $\pm 2\xi$ .

In all the results of Propositions 2.1–2.4, the total error within which the optimal value  $f^*$  is approached has additive form, and it includes the error terms coming from the  $\epsilon_k$ -subgradient bound  $\epsilon$  and the bound  $R$  on the noise magnitude. In Proposition 2.1, the total error also includes a term related to the size of the nonvanishing stepsize, while in Propositions 2.3 and 2.4, there is an additional error related to the inexact function values used in the dynamic stepsize. In the presence of persistent noise ( $R > 0$ ), the total error in approaching  $f^*$  is not zero even when  $\epsilon_k$ -subgradients are replaced by subgradients ( $\epsilon = 0$ ).

We do not know whether the bounds in Propositions 2.1–2.4 are sharp when all errors are persistent ( $\epsilon > 0$ ,  $R > 0$ , and  $\xi > 0$ ). However, we know that the bounds of Propositions 2.1 and 2.2 are sharp in the special case when  $R = 0$ . Specifically, in subsequent Example 2.1, we show that the error bound of Proposition 2.1 is sharp when  $R = 0$  and  $\epsilon = 0$ , in which case the bound depends only on the stepsize value  $\alpha$ . Furthermore, in Example 2.2, we show that the error bound of Proposition 2.2 is sharp when  $R = 0$ .

*Example 2.1* Consider the problem of minimizing  $f(x) = C^2|x|$  over  $x \in \Re$ . Consider the subgradient method using a constant step  $\alpha$  and starting with  $x_0 = \alpha C^2/2$ . The subgradient of  $f$  at  $x_0$  is  $g_0 = C^2$  and the next iterate is  $x_1 = x_0 - \alpha C^2 = -x_0$ . Subsequently, the subgradient of  $f$  at  $x_1$  is  $g_1 = -C^2$ , and the next iterate is  $x_2 = x_0$ . Therefore, the method generates a sequence  $\{x_k\}$  that “oscillates” between  $x_0$  and  $-x_0$ . The function value is constant along this sequence, i.e.,  $f(x_k) = \alpha C^2/2$  for all  $k$ . Since  $f^* = 0$ , we have

$$\liminf_{k \rightarrow \infty} f(x_k) - f^* = \frac{\alpha C^2}{2},$$

thus, showing that the error due to the use of a constant stepsize is persistent, and that the estimate of Proposition 2.1 is sharp when  $R = 0$  and  $\epsilon = 0$ .

*Example 2.2* Consider the problem of minimizing  $f(x) = |x|$  over  $x \in \mathfrak{R}$ . For  $\epsilon > 0$ , it can be seen that

$$\partial_\epsilon f(x) = \begin{cases} [-1, -1 - \frac{\epsilon}{x}] & \text{for } x < -\frac{\epsilon}{2}, \\ [-1, 1] & \text{for } x \in [-\frac{\epsilon}{2}, \frac{\epsilon}{2}], \\ [1 - \frac{\epsilon}{x}, 1] & \text{for } x > \frac{\epsilon}{2}. \end{cases}$$

Consider the  $\epsilon$ -subgradient method starting with  $x_0 = \epsilon$ . We have

$$\partial_\epsilon f(x_0) = \left[ 1 - \frac{\epsilon}{x_0}, 1 \right] = [0, 1].$$

Thus,  $g = 0$  is an  $\epsilon$ -subgradient of  $f$  at  $x_0$ . Suppose that at  $x_0$  we use the direction  $g_0 = g$ . Then, the  $\epsilon$ -subgradient method starting at  $x_0$  (with any stepsize) does not move, and converges trivially to  $x_0$ . Since  $f^* = 0$  and  $f(x_0) = x_0$ , we have

$$\liminf_{k \rightarrow \infty} f(x_k) - f^* = x_0 = \epsilon.$$

This shows that the error  $\epsilon$  is persistent, and also makes the estimate of Proposition 2.2 sharp when  $R = 0$ .

Note that Example 2.2 also makes the estimate of Proposition 2.4 sharp when  $R = 0$  and  $\xi = 0$ .

### 3 Convergence properties for $f$ with a sharp set of minima

In this section we assume that the objective function  $f$  has a linear growth property: it increases at least linearly as we move to nonoptimal feasible points starting from the set of optimal solutions. In particular, we say that a convex function  $f$  has a *sharp set of minima* over a convex set  $X$ , when the optimal set  $X^*$  is nonempty and for some scalar  $\mu > 0$  there holds

$$f(x) - f^* \geq \mu \operatorname{dist}(x, X^*), \quad \forall x \in X. \tag{3.1}$$

For such a function, we have the following result.

**Lemma 3.1** *Let the function  $f$  have a sharp set of minima. Then, for a sequence  $\{x_k\}$  generated by the method with any of the stepsize rules (a)–(c), we have for all  $k$*

$$\begin{aligned} (\operatorname{dist}(x_{k+1}, X^*))^2 &\leq (\operatorname{dist}(x_k, X^*))^2 - 2\alpha_k \frac{\mu - \|r_k\|}{\mu} (f(x_k) - f^*) \\ &\quad + 2\alpha_k \epsilon_k + \alpha_k^2 \|\tilde{g}_k\|^2. \end{aligned}$$

*Proof* The relation is implied by Lemma 2.1 and the property of  $f$  in Eq. (3.1)  $\square$

In what follows we consider a noise sequence  $\{r_k\}$  whose norm bound  $R$  is lower than  $\mu$ , i.e.,  $R < \mu$ , which we refer to as *low level noise*. In particular, we assume the following.

**Assumption 3.1** The function  $f$  has a sharp set of minima [cf. Eq. (3.1)]. The noise  $r_k$  and the errors  $\epsilon_k$  satisfy Assumption 2.2. Furthermore,  $\{r_k\}$  is a low level noise (i.e.  $R < \mu$ ).

For the constant and diminishing stepsize rules, we also assume the following.

**Assumption 3.2** There is a positive scalar  $C$  such that

$$\|g\| \leq C, \quad \forall g \in \partial_{\epsilon_k} f(x_k), \quad \forall k \geq 0,$$

where  $\partial_{\epsilon_k} f(x)$  is the set of all  $\epsilon_k$ -subgradients of  $f$  at  $x$ .

Assumptions 3.1 and 3.2 hold, for example, when the optimal set  $X^*$  is nonempty and the function  $f$  is polyhedral, i.e.,

$$f(x) = \max_{1 \leq j \leq p} \{a'_j x + b_j\},$$

where  $a_j \in \Re^n$  and  $b_j \in \Re$  for all  $j$ , in which case the scalars  $\mu$  and  $C$  are given by

$$\mu = \min_{1 \leq j \leq p} \{\|a_j\| \mid a_j \neq 0\}, \quad C = \max_{1 \leq j \leq p} \|a_j\|.$$

In the next two propositions, we give the convergence results for the method with a constant and a diminishing stepsize.

**Proposition 3.1** *Let Assumptions 3.1 and 3.2 hold. Then, for a sequence  $\{x_k\}$  generated by the method with the constant stepsize rule, we have*

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \frac{\mu}{\mu - R} \left( \epsilon + \frac{\alpha}{2} (C + R)^2 \right).$$

*Proof* The proof is based on Lemma 3.1 and a line of analysis similar to that of Proposition 2.1.  $\square$

**Proposition 3.2** *Let Assumptions 3.1 and 3.2 hold. Then, for a sequence  $\{x_k\}$  generated by the method with the diminishing stepsize rule, we have*

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \frac{\mu \epsilon}{\mu - R}.$$

*Proof* The proof uses Lemma 3.1, and a line of analysis similar to that of Proposition 2.1 combined with the fact  $\sum_{k=0}^{\infty} \alpha_k = \infty$ .  $\square$

In the convergence analysis of the dynamic stepsize rule, we use a condition on the subgradients  $g_k$  weaker than Assumption 3.2, and a condition on the parameters  $\gamma_k$  stronger than (1.6) These conditions are stated in the following.

**Assumption 3.3** The sequence  $\{g_k\}$  is bounded whenever  $\{\text{dist}(x_k, X^*)\}$  is bounded. The parameters  $\gamma_k$  in the dynamic stepsize rule (1.6) are such that

$$\gamma_k \leq 2 \frac{\mu - R}{\mu}, \quad \forall k \geq 0.$$

The first condition in Assumption 3.3 holds, for example, when  $X^*$  is bounded. Using this assumption,<sup>5</sup> we give a convergence property of the dynamic stepsize rule and the adjustment procedure (2.3)–(2.4).

**Proposition 3.3** *Let Assumptions 2.3, 3.1 and 3.3 hold. Then, for a sequence  $\{x_k\}$  generated by the method and the dynamic stepsize rule with the adjustment procedure (2.3)–(2.4), we have*

$$\inf_{k \geq 0} f(x_k) \leq f^* + 2\xi + \frac{\mu\epsilon}{\mu - R} + \delta.$$

*Proof* To arrive at a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) > f^* + 2\xi + \frac{\mu\epsilon}{\mu - R} + \delta,$$

or equivalently

$$\frac{\mu - R}{\mu} \left( \inf_{k \geq 0} f(x_k) - \delta - 2\xi - f^* \right) > \epsilon.$$

Then, in view of  $\tilde{f}(x_k) = f(x_k) + \xi_k$  and  $|\xi_k| \leq \xi$  for all  $k$ , it follows that

$$\frac{\mu - R}{\mu} \left( \inf_{k \geq 0} \tilde{f}(x_k) - \delta - \xi - f^* \right) > \epsilon. \tag{3.2}$$

By the adjustment procedure (2.3)–(2.4), we have  $\delta_k \geq \delta$  for all  $k$ . Hence, each time the target level is attained [i.e.,  $\tilde{f}(x_k) \leq \tilde{f}_{k-1}^{\text{lev}}$ ], the current best function value  $\min_{0 \leq j \leq k} \tilde{f}(x_j)$  decreases by at least  $\delta$ . Thus, in view of Eq. (3.2) the target level can be attained only a finite number of times. Then, according to Eq. (2.4), there is a nonnegative integer  $k_0$  such that

$$\delta_k = \delta, \quad \forall k \geq k_0,$$

so that the target levels  $\tilde{f}_k^{\text{lev}}$  satisfy

$$\tilde{f}_k^{\text{lev}} = \min_{0 \leq j \leq k} \tilde{f}(x_j) - \delta, \quad \forall k \geq k_0.$$

<sup>5</sup> Our subsequent results only require that the condition  $\gamma_k \leq 2 \frac{\mu - R}{\mu}$  holds for large enough  $k$ .

By choosing a larger  $k_0$  if necessary, from the preceding relation and Eq. (3.2) it can be seen that for some positive scalar  $\nu$  we have

$$\frac{\mu - R}{\mu} (\tilde{f}_k^{\text{lev}} - \xi_k - f^*) \geq \epsilon_k + \nu, \quad \forall k \geq k_0. \tag{3.3}$$

Now, by using Lemma 3.1, the relation  $f(x_k) = \tilde{f}(x_k) - \xi_k$ , and the definition of  $\alpha_k$ , we obtain for all  $k$ ,

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \frac{\mu - R}{\mu} (\tilde{f}(x_k) - \xi_k - f^*) \\ &\quad + 2\alpha_k \epsilon_k + \alpha_k \gamma_k (\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}) \\ &= (\text{dist}(x_k, X^*))^2 - \alpha_k \left( 2 \frac{\mu - R}{\mu} - \gamma_k \right) (\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}}) \\ &\quad - 2\alpha_k \left( \frac{\mu - R}{\mu} (\tilde{f}_k^{\text{lev}} - \xi_k - f^*) - \epsilon_k \right) \\ &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \left( \frac{\mu - R}{\mu} (\tilde{f}_k^{\text{lev}} - \xi_k - f^*) - \epsilon_k \right), \end{aligned} \tag{3.4}$$

where in the last inequality above we use the facts  $\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}} \geq \delta_k > 0$  and  $2(\mu - R)/\mu - \gamma_k \geq 0$  for all  $k$ . By substituting Eq. (3.3) in the preceding inequality, we have for all  $k \geq k_0$

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \nu \leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 - 2\nu \sum_{j=k_0}^k \alpha_j, \tag{3.5}$$

implying the boundedness of  $\{\text{dist}(x_k, X^*)\}$ . Hence  $\{g_k\}$  is also bounded (cf. Assumption 3.3), so that  $\|\tilde{g}_k\| \leq C + R$  for all  $k$ , where  $C$  is such that  $\|g_k\| \leq C$  for all  $k$ . Using the boundedness of  $\tilde{g}_k$  and the fact  $\tilde{f}(x_k) - \tilde{f}_k^{\text{lev}} \geq \delta$  for all  $k$ , from the definition of  $\alpha_k$  we obtain

$$\alpha_k \geq \underline{\gamma} \frac{\delta}{(C + R)^2}, \quad \forall k \geq 0,$$

which when substituted in Eq. (3.5) yields a contradiction for sufficiently large  $k$ .  $\square$

In the next proposition, we give the convergence properties of the path-based procedure.

**Proposition 3.4** *Let Assumptions 2.3, 3.1 and 3.3 hold. Then, for a sequence  $\{x_k\}$  generated by the method and the path-based procedure, we have*

$$\inf_{k \geq 0} f(x_k) \leq f^* + 2\xi + \frac{\mu\epsilon}{\mu - R}.$$



*Proof* We at first show that  $l \rightarrow \infty$ . Assume that  $l$  takes only a finite number of values, say  $l = 0, 1, \dots, \bar{l}$ , then at Step 3, we have for all  $k > k(\bar{l})$

$$B \geq \sigma_k = \sigma_{k-1} + \alpha_{k-1} \|\tilde{g}_{k-1}\| \geq \sum_{j=k(\bar{l})}^{k-1} \alpha_j \|\tilde{g}_j\| \geq \sum_{j=k(\bar{l})}^{k-1} \|x_{j+1} - x_j\|.$$

Therefore  $\alpha_k \|\tilde{g}_k\| \rightarrow 0$  and  $\{x_k\}$  is bounded, so that  $\|\tilde{g}_k\| \leq C + R$  for all  $k$ , where  $C$  is such that  $\|g_k\| \leq C$  for all  $k$ . Thus from the definition of  $\alpha_k$  we obtain

$$\alpha_k \|\tilde{g}_k\| \geq \underline{\gamma} \frac{\delta_{\bar{l}}}{2(C + R)}, \quad \forall k \geq k(\bar{l}),$$

contradicting the fact  $\alpha_k \|\tilde{g}_k\| \rightarrow 0$ . Hence, we must have  $l \rightarrow \infty$ .

Now, in order to arrive at a contradiction, assume that  $\inf_{k \geq 0} f(x_k) > f^* + 2\xi + \frac{\mu\epsilon}{\mu - R}$ , implying that

$$\inf_{k \geq 0} \tilde{f}(x_k) > f^* + \xi + \frac{\mu\epsilon}{\mu - R}. \tag{3.6}$$

If  $\delta_l$  is decreased at Step 3 only a finite number of times, then there must be an infinite number of sufficient descents, so that for some nonnegative integer  $\bar{l}$  we have  $\delta_l = \delta_{\bar{l}} > 0$  for all  $l \geq \bar{l}$ . Each time a sufficient descent is detected, the current best function value  $\min_{0 \leq j \leq k} \tilde{f}(x_j)$  is decreased by at least  $\delta_{\bar{l}}/2$ , so in view of Eq. (3.6) there can be only a finite number of sufficient descents, which is a contradiction. Hence  $\delta_l$  must be decreased at Step 3 infinitely often so that  $\lim_{l \rightarrow \infty} \delta_l = 0$ . Let  $\bar{l}$  be a sufficiently large positive integer and  $\nu$  be a positive scalar such that [cf. Eq. (3.6)]

$$\frac{\mu - R}{\mu} \left( \inf_{j \geq 0} \tilde{f}(x_j) - \delta_l - \xi - f^* \right) \geq \epsilon_k + \nu, \quad \forall k \geq k(l), \quad \forall l \geq \bar{l}.$$

Then by using the fact  $\tilde{f}_k^{\text{lev}} = \min_{0 \leq j \leq k(l)} \tilde{f}(x_j) - \delta_l$  and  $\xi \geq \xi_k$ , we obtain

$$\frac{\mu - R}{\mu} \left( \tilde{f}_k^{\text{lev}} - \xi_k - f^* \right) \geq \epsilon_k + \nu, \quad \forall k \geq k(\bar{l}).$$

Similar to the proof of Proposition 3.3, it can be seen that for all  $k$  we have [cf. Eq. (3.4)]

$$\left( \text{dist}(x_{k+1}, X^*) \right)^2 \leq \left( \text{dist}(x_k, X^*) \right)^2 - 2\alpha_k \left( \frac{\mu - R}{\mu} \left( \tilde{f}_k^{\text{lev}} - \xi_k - f^* \right) - \epsilon_k \right),$$

from which, by using the preceding relation, we obtain for all  $k \geq k(\bar{l})$

$$\begin{aligned} \left( \text{dist}(x_{k+1}, X^*) \right)^2 &\leq \left( \text{dist}(x_k, X^*) \right)^2 - 2\alpha_k \nu \\ &\leq \dots \leq \left( \text{dist}(x_{k(\bar{l})}, X^*) \right)^2 - 2\nu \sum_{j=k(\bar{l})}^k \alpha_j. \end{aligned}$$

Hence

$$\sum_{k=0}^{\infty} \alpha_k < \infty \quad (3.7)$$

and the sequence  $\{\text{dist}(x_k, X^*)\}$  is bounded. By Assumption 3.3 the sequence  $\{g_k\}$  is bounded, and therefore  $\|\tilde{g}_k\| \leq C + R$  for all  $k$ , where  $C$  is such that  $\|g_k\| \leq C$  for all  $k$ . Then, we consider the index set  $L$  given by

$$L = \left\{ l \in \{1, 2, \dots\} \mid \delta_l = \frac{\delta_{l-1}}{2} \right\},$$

and by using the same line of analysis as in the corresponding part of the proof of Proposition 2.4, we obtain the relation  $\sum_{k=0}^{\infty} \alpha_k > \infty$ , thus contradicting Eq. (3.7).  $\square$

We now discuss how the noise  $r_k$  and the  $\epsilon_k$  errors affect the established error estimate results. We consider two extreme cases: the case when  $\epsilon = 0$  and the low level noise is persistent ( $\mu > R > 0$ ), and the case when  $\epsilon > 0$  and there is no noise ( $R = 0$ ).

*When subgradients instead of  $\epsilon$ -subgradients are used (i.e.,  $\epsilon = 0$ ) and the low level noise is persistent, the error in the estimate of Proposition 3.2 vanishes and the convergence to the exact value  $f^*$  is obtained. By contrast, exact convergence cannot be guaranteed in the corresponding result of Proposition 2.2 of Sect. 2. In particular, the error estimate of Proposition 2.2 is persistent (does not diminish to zero) even when  $\epsilon = 0$ , and thus only convergence to an approximation of the value  $f^*$  can be guaranteed.*

*When approximate subgradients are used (i.e.,  $\epsilon > 0$ ) and there is no noise ( $R = 0$ ), the resulting error in the estimates of Props. 3.2 and 3.4 does not vanish even when the function evaluations are exact ( $\xi = 0$  in Proposition 3.4). In particular, the resulting error is proportional to the limiting error  $\epsilon$  associated with  $\epsilon_k$ -subgradients used in the method. This demonstrates the different nature of the noise  $r_k$  and the errors associated with using  $\epsilon_k$ -subgradients.*

We do not know whether the error bounds of Propositions 3.1–3.4 are sharp when all errors are persistent ( $\epsilon > 0$ ,  $R > 0$  and  $\xi > 0$ .) We note, however, that the estimates of Propositions 3.1–3.4 are identical to those of Propositions 2.1–2.4, respectively, when there is no noise ( $R = 0$ ). In this case, Examples 2.1 and 2.2 on sharpness of the results in Propositions 2.1 and 2.2 obviously apply to the results of Propositions 3.1 and 3.2. In addition, our estimate of Proposition 3.2 is tight when  $\epsilon > 0$  and  $1 > R \geq 0$ , as seen in the following example.<sup>6</sup>

*Example 3.1* Consider the problem of Example 2.2. The set of minima is  $X^* = \{0\}$  and the sharp minima parameter  $\mu$  is equal to 1. Let the  $\epsilon$ -subgradient noise bound be  $R$  with  $0 \leq R < 1$ . Consider the  $\epsilon$ -subdifferential  $\partial_\epsilon f(x_0)$  at a point  $x_0 = \frac{\epsilon}{1-R}$ .

<sup>6</sup> This example is based on Example 5.1 given by A. Belloni in *Lecture Notes for IAP 2005 Course*, which is available at <http://web.mit.edu/belloni>.

According to the form of the  $\epsilon$ -subdifferential set  $\partial_\epsilon f(x)$ , as given in Example 2.2, we have

$$\partial_\epsilon f(x_0) = \left[ 1 - \frac{\epsilon}{x_0}, 1 \right] = [R, 1].$$

Thus,  $g = R$  is an  $\epsilon$ -subgradient of  $f$  at  $x_0$ . Suppose that at  $x_0$  we use the noisy direction  $\tilde{g}_0 = g + r$  with the noise  $r = -R$ . Then  $\tilde{g}_0 = 0$ , the method (1.2) starting at  $x_0$  does not move, and converges trivially to  $x_0$ . Since  $f^* = 0$  and  $f(x_0) = x_0$ , we have

$$\liminf_{k \rightarrow \infty} f(x_k) - f^* = x_0 = \frac{\epsilon}{1 - R}.$$

This shows that the error  $\epsilon$  is persistent, and also makes the estimate of Proposition 3.2 sharp (for  $\mu = 1$ ).

#### 4 Implications for incremental $\epsilon_k$ -subgradient methods

In this section we consider a special case of problem (1.1), where the function  $f$  is the sum of a large number of component functions  $f_i$ , i.e.,

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

with each  $f_i : \mathfrak{N}^n \rightarrow \mathfrak{N}$  convex. In this case, to solve the problem, an incremental method can be applied, which exploits the special structure of  $f$ . The incremental method is similar to the method (1.2). The main difference is that at each iteration,  $x$  is changed incrementally through a sequence of  $m$  steps. Each step is a noisy subgradient iteration for a single component function  $f_i$ , and there is one step per component function. Thus, an iteration can be viewed as a cycle of  $m$  subiterations. If  $x_k$  is the vector obtained after  $k$  cycles, the vector  $x_{k+1}$  obtained after one more cycle is

$$x_{k+1} = \psi_{m,k}, \tag{4.1}$$

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

$$\psi_{i,k} = \mathcal{P}_X [\psi_{i-1,k} - \alpha_k \tilde{g}_{i,k}], \quad i = 1, \dots, m, \tag{4.2}$$

with

$$\tilde{g}_{i,k} = g_{i,k} + r_{i,k}, \tag{4.3}$$

where  $g_{i,k}$  is an  $\epsilon_{i,k}$ -subgradient of  $f_i$  at  $\psi_{i-1,k}$ ,  $r_{i,k}$  is a noise, and

$$\psi_{0,k} = x_k. \tag{4.4}$$

Without the presence of noise, the incremental method has been studied by Kibardin [14], Nedić and Bertsekas [18, 19], Nedić, Bertsekas, and Borkar [17] (see also Nedić [20]), Ben-Tal, Margalit, and Nemirovski [1], and Kiwiel [15]. The incremental idea have been extended to min-max problems through the use of bundle methods by Gaudioso, Giallombardo, and Miglionico [11]. The presence of noise in incremental subgradient methods was addressed by Solodov and Zavriev in [26] for a compact constraint set  $X$  and the diminishing stepsize rule. See [3] for an extensive reference on incremental subgradient methods.

#### 4.1 Convergence results for a compact constraint set

Here, we show that the results of Sect. 2 apply to incremental method (4.1)–(4.4). For this, we use the following boundedness assumption on the noise  $r_{i,k}$  and the  $\epsilon_{i,k}$ -subgradients.

**Assumption 4.1** There exist positive scalars  $R_1, \dots, R_m$  such that for each  $i = 1, \dots, m$ ,

$$\|r_{i,k}\| \leq R_i \quad \forall k \geq 0. \tag{4.5}$$

There exist scalars  $\epsilon_1 \geq 0, \dots, \epsilon_m \geq 0$  such that for each  $i = 1, \dots, m$ ,

$$\limsup_{k \rightarrow \infty} \epsilon_{ik} = \epsilon_i.$$

Under the assumption  $\limsup_{k \rightarrow \infty} \epsilon_{ik} = \epsilon_i$  for some scalar  $\epsilon_i \geq 0$ , it can be seen that  $\sup_k \epsilon_{ik}$  is finite. Let us denote it by  $\tilde{\epsilon}_i$ , i.e., for each  $i = 1, \dots, m$ ,

$$\tilde{\epsilon}_i = \sup_k \epsilon_{ik}.$$

Since the  $\epsilon$ -subdifferential sets are nested as  $\epsilon$  increases, in view of  $\epsilon_{i,k} \leq \tilde{\epsilon}_i$ , it follows that  $\partial_{\epsilon_{i,k}} f_i(x) \subseteq \partial_{\tilde{\epsilon}_i} f_i(x)$  for any  $x$ . Therefore, for each  $i$

$$\cup_k \partial_{\epsilon_{i,k}} f_i(\psi_{i-1,k}) \subseteq \cup_k \partial_{\tilde{\epsilon}_i} f_i(\psi_{i-1,k}) \subseteq \cup_{x \in X} \partial_{\tilde{\epsilon}_i} f_i(x).$$

Under the compactness of the set  $X$ , the set  $\cup_{x \in X} \partial_{\tilde{\epsilon}_i} f_i(x)$  is compact for each  $i$  (see Dem’yanov and Vasil’ev [6], Corollary on page 77), implying that there exist a constant  $C_i$  such that for all  $x \in X$ ,

$$\|g\| \leq C_i \quad \forall g \in \partial_{\epsilon_{i,k}} f_i(x) \quad \text{and} \quad \forall k \geq 0. \tag{4.6}$$

Since the subdifferential set  $\partial f_i(x)$  is contained in the  $\epsilon$ -subdifferential set  $\partial_{\epsilon_{i,k}} f_i(x)$ , it follows that for all  $i = 1, \dots, m$  and  $x \in X$ ,

$$\|g\| \leq C_i \quad \forall g \in \partial f_i(x). \tag{4.7}$$

We now give a basic lemma which will be used for the analysis of the incremental methods similar to the manner in which Lemma 2.1 was used for the non-incremental methods. For a compact set  $X$  (cf. Assumption 2.1), we have the following result.

**Lemma 4.1** *Let Assumptions 2.1 and 4.1 hold. Then, for a sequence  $\{x_k\}$  generated by the incremental method and any stepsize rule, we have for all  $k$*

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (f(x_k) - f^*) + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k \tilde{R} \text{dist}(x_k, X^*) + \alpha_k^2 ((\tilde{C} + \tilde{R})^2 - \tilde{S}), \end{aligned}$$

where  $\epsilon_k = \sum_{i=1}^m \epsilon_{i,k}$  for all  $k$ , and

$$\tilde{R} = \sum_{i=1}^m R_i, \quad \tilde{C} = \sum_{i=1}^m C_i, \quad \tilde{S} = 2 \sum_{i=1}^{m-1} R_i \sum_{j=i+1}^m C_j + 2 \sum_{i=2}^m R_i \sum_{j=1}^{i-1} R_j. \quad (4.8)$$

*Proof* Using the nonexpansion property of the projection, the noise and the subdifferential boundedness [cf. Eqs. (4.5) and (4.6)], we obtain for all  $y \in X$ , all  $i$ , and all  $k$ ,

$$\begin{aligned} \|\psi_{i,k} - y\|^2 &\leq \|\psi_{i-1,k} - \alpha_k \tilde{g}_{i,k} - y\|^2 \\ &\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k \tilde{g}'_{i,k}(\psi_{i-1,k} - y) + \alpha_k^2 \|\tilde{g}_{i,k}\|^2 \\ &\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k (f_i(\psi_{i-1,k}) - f_i(y)) + 2\alpha_k \epsilon_{i,k} \\ &\quad + 2\alpha_k R_i \|\psi_{i-1,k} - y\| + \alpha_k^2 (C_i + R_i)^2, \end{aligned} \quad (4.9)$$

where the last inequality follows from the fact

$$\begin{aligned} \tilde{g}'_{i,k}(\psi_{i-1,k} - y) &= g'_{i,k}(\psi_{i-1,k} - y) + r'_{i,k}(\psi_{i-1,k} - y) \\ &\geq g'_{i,k}(\psi_{i-1,k} - y) - R_i \|\psi_{i-1,k} - y\| \end{aligned}$$

[cf. Eqs. (4.3) and (4.5)] and the  $\epsilon_{i,k}$ -subgradient inequality for  $f_i$  at  $\psi_{i-1,k}$

$$g'_{i,k}(\psi_{i-1,k} - y) \geq f_i(\psi_{i-1,k}) - f_i(y) - \epsilon_{i,k}, \quad \forall y \in \mathfrak{R}^n.$$

By summing over  $i$  in (4.9) and by using  $\epsilon_k = \sum_{i=1}^m \epsilon_{i,k}$ , we have for all  $y \in X$  and  $k$

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k \sum_{i=1}^m (f_i(\psi_{i-1,k}) - f_i(y)) + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k \sum_{i=1}^m R_i \|\psi_{i-1,k} - y\| + \alpha_k^2 \sum_{i=1}^m (C_i + R_i)^2 \end{aligned}$$

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

where  $\tilde{R} = \sum_{i=1}^m R_i$ . By using the subdifferential boundedness and the fact  $\|\psi_{i,k} - x_k\| \leq \alpha_k \sum_{j=1}^i C_j$  for all  $i, k$ , from the preceding relation we obtain

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (f(x_k) - f(y)) + 2\alpha_k^2 \sum_{i=2}^m C_i \sum_{j=1}^{i-1} C_j + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k \tilde{R} \|x_k - y\| + 2\alpha_k^2 \sum_{i=2}^m R_i \sum_{j=1}^{i-1} C_j + \alpha_k^2 \sum_{i=1}^m (C_i + R_i)^2 \\ &= \|x_k - y\|^2 - 2\alpha_k (f(x_k) - f(y)) + 2\alpha_k \epsilon_k + 2\alpha_k \tilde{R} \|x_k - y\| \\ &\quad + \alpha_k^2 \left( 2 \sum_{i=2}^m (C_i + R_i) \sum_{j=1}^{i-1} C_j + \sum_{i=1}^m (C_i + R_i)^2 \right). \end{aligned}$$

After some calculation, it can be seen that

$$\begin{aligned} 2 \sum_{i=2}^m (C_i + R_i) \sum_{j=1}^{i-1} C_j + \sum_{i=1}^m (C_i + R_i)^2 &= (\tilde{C} + \tilde{R})^2 \\ - 2 \sum_{i=1}^{m-1} R_i \sum_{j=i+1}^m C_j - 2 \sum_{i=2}^m R_i \sum_{j=1}^{i-1} R_j &= (\tilde{C} + \tilde{R})^2 - \tilde{S}, \end{aligned}$$

where  $\tilde{C}$ ,  $\tilde{R}$  and  $\tilde{S}$  are as given in Eq. 4.8. From the preceding two relations we obtain for all  $y \in X$  and  $k$

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (f(x_k) - f(y)) + 2\alpha_k \epsilon_k \\ &\quad + 2\alpha_k \tilde{R} \|x_k - y\| + \alpha_k^2 ((\tilde{C} + \tilde{R})^2 - \tilde{S}), \end{aligned}$$

The desired inequality follows from the preceding relation, by letting  $y = \mathcal{P}_{X^*}[x_k]$  and using the relations

$$\|x_k - \mathcal{P}_{X^*}[x_k]\| = \text{dist}(x_k, X^*), \quad \text{dist}(x_{k+1}, X^*) \leq \|x_{k+1} - \mathcal{P}_{X^*}[x_k]\|.$$

□

We consider the incremental method using the constant, the diminishing, and the modified dynamic stepsize rule. The modification of the dynamic stepsize rules consists of replacing  $\|\tilde{g}_k\|^2$  by  $(\tilde{C} + \tilde{R})^2 - \tilde{S}$  in Eq. (1.6), and at Step 5 of the path-based

procedure, the parameter  $\sigma_k$  should be updated by

$$\sigma_{k+1} = \sigma_k + \alpha_k \sqrt{(\tilde{C} + \tilde{R})^2 - \tilde{S}}.$$

In this case, however, the modified dynamic stepsize rule may result in a small stepsize value  $\alpha_k$ .

Using Assumptions 2.1 and 4.1, we can show that the results of Props. 2.1–2.4 apply to a sequence  $\{x_k\}$  generated by the incremental method, where in the estimates of Sect. 2 we replace  $R$  by  $\tilde{R}$  and  $(C + R)^2$  by  $(\tilde{C} + \tilde{R})^2 - \tilde{S}$ , and we let  $\epsilon = \sum_{i=1}^m \epsilon_i$  (with  $\epsilon_i$  as defined in Assumption 4.1). This can be seen by using Lemma 4.1 in place of Lemma 2.1.

### 4.2 Convergence results for $f$ with sharp set of minima

In this section, we show that the results of Sect. 3 also hold for an incremental  $\epsilon_k$ -subgradient method. We consider an objective function  $f$  with a sharp set of minima, as defined in Eq. (3.1).

We also use the following subgradient boundedness assumption.

**Assumption 4.2** There exist scalars  $C_1, \dots, C_m$  such that for each  $i = 1, \dots, m$ ,

$$\|g\| \leq C_i \quad \forall g \in \partial f_i(x_k) \cup \partial_{\epsilon_{i,k}} f_i(\psi_{i-1,k}) \quad \text{and} \quad \forall k \geq 0,$$

where  $\partial f_i(x)$  and  $\partial_{\epsilon} f_i(x)$  denote the sets of all subgradients and  $\epsilon$ -subgradients of  $f_i$  at  $x$ , respectively.

This assumption holds, for example, when each function  $f_i$  is polyhedral. Under the two preceding assumptions, we have a refinement of the basic relation shown in Lemma 4.1, as follows.

**Lemma 4.2** *Let Assumptions 4.1 and 4.2 hold. Assume also that the function  $f$  has a sharp set of minima [cf. Eq. (3.1)]. Then, for a sequence  $\{x_k\}$  generated by the incremental method and any stepsize rule, we have for all  $k$*

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \frac{\mu - \tilde{R}}{\mu} (f(x_k) - f^*) \\ &\quad + 2\alpha_k \epsilon_k + \alpha_k^2 ((\tilde{C} + \tilde{R})^2 - \tilde{S}), \end{aligned}$$

where  $\epsilon_k = \sum_{i=1}^m \epsilon_{i,k}$  for all  $k$ , and the scalars  $\tilde{R}$ ,  $\tilde{C}$ , and  $\tilde{S}$  are given in Eq. (4.8).

*Proof* Similar to the proof of Lemma 4.1, using Assumption 4.1 and the subgradient boundedness of Assumption 4.2, we can show that the basic relation of Lemma 4.1 holds. In particular, we have for all  $k$

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k (f(x_k) - f^*) + 2\alpha_k \epsilon_k + 2\alpha_k \tilde{R} \text{dist}(x_k, X^*) \\ &\quad + \alpha_k^2 ((\tilde{C} + \tilde{R})^2 - \tilde{S}). \end{aligned}$$

Since  $f$  has a sharp set of minima, it follows by Eq. (3.1) that  $\text{dist}(x_k, X^*) \leq (f(x_k) - f^*)/\mu$ . By substituting this relation in the preceding inequality, we immediately obtain the desired relation.  $\square$

For the incremental method, the noise  $r_{i,k}$  is a low level noise when  $\tilde{R} < \mu$  where  $\tilde{R} = \sum_{i=1}^m R_i$  and  $R_i$  is the norm bound on the noise sequence  $\{r_{i,k}\}$  as in Assumption 4.1. For a function  $f$  with sharp minima and low level noise, using Assumptions 4.1 and 4.2, we can show that the results of Props. 3.1–3.4 apply to a sequence  $\{x_k\}$  generated by the incremental method. In this case, the results of Sect. 3 hold with  $\tilde{R}$  instead of  $R$  and with  $\epsilon = \sum_{i=1}^m \epsilon_i$ , where  $\epsilon_i$  is as given in Assumption 4.1. (In case of Proposition 3.1, we also have  $(\tilde{C} + \tilde{R})^2 - \tilde{S}$  instead of  $(C + R)^2$ .) This can be seen by using Lemma 4.2 in place of Lemma 3.1 and a line of analysis identical to that of Sect. 3.

## References

1. Ben-Tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization method and its use for the positron emission tomography reconstruction. *SIAM J. Optim.* **12**(1), 79–108 (2001)
2. Bertsekas, D.P.: *Nonlinear programming*, 2nd edn. Athena Scientific, Belmont (1999)
3. Bertsekas, D.P., Nedić, A., Ozdaglar, A.E.: *Convex Analysis and Optimization*. Athena Scientific, Belmont (2003)
4. Brännlund, U.: *On relaxation methods for nonsmooth convex optimization*. Doctoral Thesis, Royal Institute of Technology, Stockholm (1993)
5. Burke, J.V., Ferris, M.C.: Weak sharp minima in mathematical programming. *SIAM J. Control Optim.* **31**(5), 1340–1359 (1993)
6. Dem'yanov, V.F., Vasil'ev, L.V.: *Nondifferentiable Optimization*. Optimization Software Inc., New York (1985)
7. Ermoliev, Yu.M.: On the stochastic quasi-gradient method and stochastic quasi-feyer sequences. *Kibernetika* **2**, 73–83 (1969)
8. Ermoliev, Yu.M.: *Stochastic Programming Methods*. Nauka, Moscow (1976)
9. Ermoliev, Yu.M.: Stochastic quasigradient methods and their application to system optimization. *Stochastics* **9**, 1–36 (1983)
10. Ermoliev, Yu.M.: Stochastic quasigradient methods. In: Ermoliev, Yu.M., Wets, R.J.-B. (eds.) *Numerical Techniques for Stochastic Optimization*. IIASA, Springer, Heidelberg, pp. 141–185 (1988)
11. Gaudio, M., Giallombardo, G., Miglionico, G.: An incremental method for solving convex finite min–max problems. *Math. Oper. Res.* **31**(1), 173–187 (2006)
12. Goffin, J.L., Kiwiel, K.: Convergence of a simple subgradient level method. *Math. Program.* **85**, 207–211 (1999)
13. Kashyap, A., Basar, T., Srikant, R.: Quantized consensus. *Automatica* **43**, 1192–1203 (2007)
14. Kibardin, V.M.: Decomposition into functions in the minimization problem. *Autom. Remote Control* **40**, 1311–1323 (1980)
15. Kiwiel, K.C.: Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM J. Optim.* **14**(3), 807–840 (2004)
16. Kiwiel, K.C.: A proximal bundle method with approximate subgradient linearizations. *SIAM J. Optim.* **16**(4), 1007–1023 (2006)
17. Nedić, A., Bertsekas, D.P., Borkar, V.: Distributed asynchronous incremental subgradient methods. In: Butnariu, D., Censor Y., Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*. Stud. Comput. Math., Elsevier, Amsterdam (2001)
18. Nedić, A., Bertsekas, D.P.: Convergence rate of incremental subgradient algorithm. In: Uryasev S., Pardalos P.M. (eds.) *Stochastic Optimization: Algorithms and Applications*. Kluwer, Dordrecht, pp. 263–304 (2000)
19. Nedić, A., Bertsekas, D.P.: Incremental subgradient methods for nondifferentiable optimization. *SIAM J. Optim.* **12**, 109–138 (2001)



20. Nedić, A.: Subgradient Methods for Convex Optimization. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2002)
21. Nurminskii, E.A.: Minimization of nondifferentiable functions in presence of noise. *Kibernetika* **10**(4), 59–61 (1974)
22. Pang, J.-S.: Error bounds in mathematical programming. *Math. Program. Ser. B* **79**, 299–332 (1997)
23. Polyak, B.T.: Nonlinear programming methods in the presence of noise. *Math. Program.* **1**(4), 87–97 (1978)
24. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc., New York (1987)
25. Rabbat, M.G., Nowak, R.D.: Quantized incremental algorithms for distributed optimization. *IEEE J. Select. Areas Commun.* **23**(4), 798–808 (2005)
26. Solodov, M.V., Zavriev, S.K.: Error stability properties of generalized gradient-type algorithms. *J. Opt. Theory Appl.* **98**(3), 663–680 (1998)
27. Tuncer, C.A., Coates, M.J., Rabbat, M.G.: Distributed average consensus with dithered quantization. *IEEE Trans. Signal Process.* **56**(10), 4905–4918 (2008)