

THE EFFECT OF DETERMINISTIC NOISE¹ IN SUBGRADIENT METHODS

by

Angelia Nedić² and Dimitri P. Bertsekas³

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. 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 subgradient errors. In the first case, the tolerance is nonzero, but in the second case, somewhat surprisingly, 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.

¹ Research supported by NSF under Grant ACI-9873339.

² Dept. of Industrial and Enterprise Systems Engineering, UIUC, Urbana, IL 61801.

³ Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA 02139.

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 ϵ -subgradient method where the ϵ -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 α_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 ϵ_k -subgradient of f at x_k for some $\epsilon_k \geq 0$, i.e., g_k satisfies

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

We quantify the joint effect of the noise level ($\max_k \|r_k\|$) and the approximate-subgradient error level ($\limsup_k \epsilon_k$), and study the convergence properties of the method (1.2) using the following stepsize rules:

(a) *Constant Stepsize Rule.* The stepsize α_k is fixed to a positive scalar α .

(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 with Known f^* .* The stepsize α_k is given by

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{\|\tilde{g}_k\|^2}, \quad 0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < 2, \quad \forall k \geq 0. \tag{1.5}$$

(d) *Dynamic Stepsize Rule with Unknown f^* .* The stepsize α_k is obtained from Eq. (1.5) by replacing the exact value f^* with an approximation f_k^{lev} of f^* , so that

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

For this stepsize rule, we consider two procedures for adjusting the target levels f_k^{lev} (cf. Section 2).

The issue of noise in the context of subgradient optimization was first studied by Ermoliev in [Erm69] (see also Ermoliev [Erm76], [Erm83], and [Erm88], and Nurminkii [Nur74]), 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 α_k , and the diminishing noise r_k and zero ϵ_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 α_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 [Pol78] and [Pol87], 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 [Pol78], or Theorem 1 in Section 5 of Chapter 5 in [Pol87]). 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 α_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 [SoZ98], 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.

In contrast with the existing literature, in this paper we are primarily concerned with cases where the noise and subgradient approximation errors are persistent (nondiminishing). We explore the noise and ϵ -subgradient error effects on convergence of subgradient methods using the stepsize rules (a)–(d), as listed above. We establish error bounds on the approximate solutions produced by the algorithms in the limit as the number of iterations increases to infinity. We quantify these error bounds explicitly in terms of the noise magnitude, ϵ -subgradient errors, and stepsize parameters. One contribution of this work is in the establishment of error bounds for noisy ϵ -subgradient methods for stepsize rules (a)–(d). Another contribution is in the use of constant and dynamic stepsize rules in noisy subgradient optimization. While these stepsize rules have been used in ϵ_k -subgradient methods (the case of $r_k \equiv 0$ and $\epsilon_k > 0$), they have not been considered even for noisy subgradient methods (the case of $r_k > 0$ and $\epsilon_k \equiv 0$).

This paper is organized as follows: In Section 2, we give the convergence properties of the method for a compact constraint set X .¹ In Section 3, we discuss the convergence properties of the method for the case

¹ The results of Section 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

when the objective function f has a set of sharp minima (also known as weak sharp minima [BuF93]). As a special case, our results show that with $\epsilon_k \equiv 0$ and the stepsize rules (a)-(d), 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).² In Section 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 Sections 2 and 3 for incremental subgradient methods, and we compare the corresponding errors.

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 Section 1.

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

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

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

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

where in the last inequality we use the ϵ_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

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

Q.E.D.

compactness of X is when f is polyhedral, and X^* is nonempty and bounded. The case of polyhedral f , however, is treated separately in Section 3, so for simplicity, in Section 2 we assume that X is compact.

² Our result for diminishing stepsize has been established by Solodov and Zavriev in [SoZ98], Lemma 4.3, under the additional assumption that the constraint set is compact.

Throughout this section, we consider the case where X is a compact set. Furthermore, we assume that the noise magnitude is bounded and that the ϵ -subgradient errors are asymptotically bounded. Specifically, we use the following assumptions.

Assumption 2.1: The constraint set X is compact.

Assumption 2.2: The noise r_k and the errors ϵ_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

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

We now give the convergence properties for each of the stepsize rules described in Section 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 ν 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

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

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

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

which yields a contradiction for sufficiently large k . **Q.E.D.**

As suggested by Prop. 2.1, we expect that the error term involving the stepsize α 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 Prop. 2.1. **Q.E.D.**

In the next proposition, we give a convergence property for the dynamic stepsize rule with known f^* .

Proposition 2.3: Let Assumptions 2.1 and 2.2 hold. Then, for a sequence $\{x_k\}$ generated by the method and the dynamic stepsize rule with known f^* , we have

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \frac{2(\epsilon + Rd)}{2 - \bar{\gamma}}.$$

Proof: To arrive at a contradiction, assume that

$$(2 - \bar{\gamma}) \liminf_{k \rightarrow \infty} (f(x_k) - f^*) > 2\epsilon + 2Rd,$$

so that for some nonnegative integer k_0 and a positive scalar ν we have

$$(2 - \bar{\gamma})(f(x_k) - f^*) \geq 2\epsilon_k + 2Rd + \nu, \quad \forall k \geq k_0. \quad (2.3)$$

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

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k(f(x_k) - f^*) + 2\alpha_k\epsilon_k + 2\alpha_kRd + \alpha_k^2\|\tilde{g}_k\|^2 \\ &\leq (\text{dist}(x_k, X^*))^2 - \alpha_k((2 - \bar{\gamma})(f(x_k) - f^*) - 2\epsilon_k - 2Rd). \end{aligned} \quad (2.4)$$

The last inequality in the preceding relation is obtained by using the definition of α_k as follows:

$$-2\alpha_k(f(x_k) - f^*) + \alpha_k^2\|\tilde{g}_k\|^2 = -\alpha_k(2 - \gamma_k)(f(x_k) - f^*) \leq -\alpha_k(2 - \bar{\gamma})(f(x_k) - f^*).$$

By using relation (2.3) in Eq. (2.4), we obtain

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - \alpha_k\nu \leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 - \nu \sum_{j=k_0}^k \alpha_j, \quad \forall k \geq k_0. \quad (2.5)$$

Furthermore, from the definition of α_k , relation (2.3), and the fact $\|\tilde{g}_k\| \leq C + R$ [cf. Eq. (2.2)] we have

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

which when substituted in Eq. (2.5) yields a contradiction for sufficiently large k . **Q.E.D.**

Next we consider the dynamic stepsize rule with unknown f^* , where the target level f_k^{lev} is given by

$$f_k^{\text{lev}} = \min_{0 \leq j \leq k} f(x_j) - \delta_k, \quad (2.6)$$

and δ_k is a positive scalar. The value δ_{k+1} is calculated as follows

$$\delta_{k+1} = \begin{cases} \bar{\beta}\delta_k & \text{if } f(x_{k+1}) \leq f_k^{\text{lev}}, \\ \max\{\underline{\beta}\delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\text{lev}}, \end{cases} \quad (2.7)$$

where $\bar{\beta}$, $\underline{\beta}$, δ_0 , and δ are fixed positive scalars, with $\bar{\beta} \geq 1$ and $\underline{\beta} < 1$. If in this procedure we set $\delta_0 = \delta$ and $\bar{\beta} = 1$, then $\delta_k = \delta$ for all k . Therefore this procedure includes, as a special case, a procedure where δ_k is fixed to a positive constant.

For a compact X , the procedure (2.6)–(2.7) gives a nonvanishing stepsize α_k , so that the convergence property of this procedure is similar to that of a constant stepsize, as shown in the following proposition.

Proposition 2.4: Let Assumptions 2.1 and 2.2 hold. Then, for a sequence $\{x_k\}$ generated by the method and the dynamic stepsize rule with unknown f^* using the adjustment procedure (2.6)–(2.7), we have

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

Proof: To arrive at a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) > f^* + \epsilon + Rd + \delta. \quad (2.8)$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\text{lev}}$], the best current function value $\min_{0 \leq j \leq k} f(x_j)$ is decreased by at least δ [cf. Eqs. (2.6)–(2.7)], so that in view of Eq. (2.8), the target level can be attained only a finite number of times. From Eq. (2.7) it follows that after finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is a nonnegative integer k_0 such that

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

By using the fact $f_k^{\text{lev}} = \min_{0 \leq j \leq k} f(x_j) - \delta$ for $k \geq k_0$ and by choosing a larger k_0 if necessary, from Eq. (2.8) it can be seen that for some positive scalar ν we have

$$f_k^{\text{lev}} - f^* \geq \epsilon_k + Rd + \nu, \quad \forall k \geq k_0. \quad (2.9)$$

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

$$\begin{aligned}
 (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k(f(x_k) - f^*) + 2\alpha_k\epsilon_k + 2\alpha_k R d + \alpha_k^2 \|\tilde{g}_k\|^2 \\
 &= (\text{dist}(x_k, X^*))^2 - 2\alpha_k(f(x_k) - f_k^{\text{lev}}) - 2\alpha_k(f_k^{\text{lev}} - f^*) + 2\alpha_k\epsilon_k \\
 &\quad + 2\alpha_k R d + \alpha_k^2 \|\tilde{g}_k\|^2 \\
 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k(f_k^{\text{lev}} - f^* - \epsilon_k - R d).
 \end{aligned} \tag{2.10}$$

The last inequality in the preceding relation follows from the definition of α_k and the following relation

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

By using inequality (2.9) in relation (2.10), we have

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \nu \leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 - 2\nu \sum_{j=k_0}^k \alpha_j, \quad \forall k \geq k_0. \tag{2.11}$$

Since $f(x_k) - f_k^{\text{lev}} \geq \delta$ and $\|\tilde{g}_k\| \leq C + R$ for all k [cf. Eq. (2.2)], from the definition of α_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. (2.11) yields a contradiction for sufficiently large k . **Q.E.D.**

In the following algorithm we describe a path-based procedure for adjusting the target levels f_k^{lev} . This procedure is based on the algorithm of Brännlund [Brä93], which was further developed by Goffin and Kiwiel [GoK99].

The Path-Based Procedure

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

Step 1 (Function evaluation) Calculate $f(x_k)$ and \tilde{g}_k , where \tilde{g}_k is given by Eq. (1.3). If $f(x_k) < f_{k-1}^{\text{rec}}$, then set $f_k^{\text{rec}} = f(x_k)$. Otherwise set $f_k^{\text{rec}} = f_{k-1}^{\text{rec}}$ [so that f_k^{rec} keeps the record of the smallest value attained by the iterates that are generated so far, i.e., $f_k^{\text{rec}} = \min_{0 \leq j \leq k} f(x_j)$].

Step 2 (Sufficient descent) If $f(x_k) \leq f_{k(l)}^{\text{rec}} - \frac{\delta_l}{2}$, 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 $f_k^{\text{lev}} = f_{k(l)}^{\text{rec}} - \delta_l$. 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 $f_k^{\text{lev}} = f_{k(l)}^{\text{rec}} - \delta_l$ 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 σ_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 f_k^{lev} is too low (i.e., sufficient descent cannot occur), then due to oscillations of x_k the parameter σ_k eventually exceeds the prescribed upper bound B on the path length and the parameter δ_l is decreased.

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

Proposition 2.5: Let Assumptions 2.1 and 2.2 hold. Then, for a sequence generated by the method and the path-based procedure, we have

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

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

$$\inf_{k \geq 0} f(x_k) > f^* + \epsilon + Rd. \quad (2.12)$$

If 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

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

Hence $l \rightarrow \infty$. If δ_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 $\min_{0 \leq j \leq k} f(x_j)$ is decreased by at least $\delta_{\bar{l}}/2$, so in view of Eq. (2.12) there can be only a finite number of sufficient descents, which is a contradiction. Therefore δ_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 ν we have [cf. Eq. (2.12)]

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

Consequently [since $f_k^{\text{lev}} = \min_{0 \leq j \leq k(l)} f(x_j) - \delta_l$]

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

Similar to the proof of Prop. 2.4, it can be seen that for all k we have [cf. Eq. (2.10)]

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

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

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

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

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} \|\tilde{g}_{k-1}\| = \sum_{j=k(l)}^{k-1} \alpha_j \|\tilde{g}_j\|, \quad (2.13)$$

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 $\sum_{k=0}^{\infty} \alpha_k$. **Q.E.D.**

Note that in the results of Props. 2.1–2.5, the total error within which the optimal value f^* is approached has additive form, and it includes the error terms coming from the ϵ_k -subgradient bound ϵ and the bound R on the noise magnitude; [in Props. 2.1 and 2.4, the total error also includes a term related to the size of the nonvanishing steplength α_k]. In the presence of persistent noise ($R > 0$), the total error in approaching f^* is not zero even when ϵ_k -subgradients are replaced by subgradients ($\epsilon = 0$).

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

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

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 and any of the stepsize rules (a)-(d), we have for all k

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

Proof: The relation is implied by Lemma 2.1 and the property of f in Eq. (3.1). **Q.E.D.**

In what follows we consider a noise sequence $\{r_k\}$ whose norm bound R is lower than μ , 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 ϵ_k satisfy Assumption 2.1. 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 ϵ_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 $a_j \in \mathfrak{R}^n$ and $b_j \in \mathfrak{R}$ for all j , in which case the scalars μ 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 Prop. 2.1. **Q.E.D.**

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 Prop. 2.1 combined with the fact $\sum_{k=0}^{\infty} \alpha_k = \infty$. **Q.E.D.**

In the convergence analysis of the dynamic stepsize rules, we can use a weaker assumption than Assumption 3.2. In particular, we assume the following.

Assumption 3.3: The sequence $\{g_k\}$ is bounded whenever $\{\text{dist}(x_k, X^*)\}$ is bounded.

The assumption holds, for example, if X^* is bounded. Using this assumption, we give a convergence property of the dynamic stepsize rule with known f^* .

Proposition 3.3: Let Assumptions 3.1 and 3.3 hold. Furthermore, let the parameter $\bar{\gamma}$ in the dynamic stepsize rule with known f^* be such that

$$2 \frac{\mu - R}{\mu} - \bar{\gamma} > 0.$$

Then, for a sequence $\{x_k\}$ generated by the method and the dynamic stepsize with known f^* , we have

$$\liminf_{k \rightarrow \infty} f(x_k) \leq f^* + \frac{2\mu\epsilon}{2(\mu - R) - \mu\bar{\gamma}}.$$

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

$$\liminf_{k \rightarrow \infty} f(x_k) > f^* + \frac{2\mu\epsilon}{2(\mu - R) - \mu\bar{\gamma}},$$

or equivalently

$$\left(2 \frac{\mu - R}{\mu} - \bar{\gamma}\right) \liminf_{k \rightarrow \infty} (f(x_k) - f^*) > 2\epsilon,$$

so that for some positive scalar ν and a nonnegative integer k_0 we have

$$\left(2 \frac{\mu - R}{\mu} - \bar{\gamma}\right) (f(x_k) - f^*) \geq 2\epsilon_k + \nu, \quad \forall k \geq k_0. \quad (3.2)$$

Based on Lemma 3.1 and the definition of the stepsize, we have

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \frac{\mu - R}{\mu} (f(x_k) - f^*) + 2\alpha_k \epsilon_k + \alpha_k \gamma_k (f(x_k) - f^*) \\ &\leq (\text{dist}(x_k, X^*))^2 - \alpha_k \left(\left(2 \frac{\mu - R}{\mu} - \bar{\gamma}\right) (f(x_k) - f^*) - 2\epsilon_k \right), \end{aligned}$$

which when combined with Eq. (3.2) implies that for all $k \geq k_0$

$$\left(\text{dist}(x_{k+1}, X^*)\right)^2 \leq \left(\text{dist}(x_k, X^*)\right)^2 - \alpha_k \nu \leq \dots \leq \left(\text{dist}(x_{k_0}, X^*)\right)^2 - \nu \sum_{j=k_0}^k \alpha_j. \quad (3.3)$$

Hence $\{\text{dist}(x_k, X^*)\}$ is bounded and, according to Assumption 3.3, so is $\{g_k\}$. Therefore $\|g_k\| \leq C$ for all k and some scalar C . Since $\|r_k\| \leq R$ [cf. Assumption 3.1], it follows that $\|\tilde{g}_k\| \leq C + R$ for all k . The boundedness of \tilde{g}_k and the definition of α_k imply that

$$\alpha_k \geq \underline{\gamma} \frac{f(x_k) - f^*}{(C + R)^2}, \quad \forall k \geq 0.$$

Because the sum $\sum_{k=0}^{\infty} \alpha_k$ is finite [cf. Eq. (3.3)], from the preceding relation we obtain $f(x_k) \rightarrow f^*$, thus contradicting Eq. (3.2). **Q.E.D.**

We now give the convergence properties of the dynamic stepsize rule using the adjustment procedure (2.6)–(2.7).

Proposition 3.4: Let Assumptions 3.1 and 3.3 hold. Furthermore, let the parameters γ_k in the dynamic stepsize rule with unknown f^* be such that

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

Then, for a sequence $\{x_k\}$ generated by the method and the dynamic stepsize rule for unknown f^* that uses the adjustment procedure (2.6)–(2.7), we have

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

Proof: To arrive at a contradiction, assume that

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

or equivalently

$$\frac{\mu - R}{\mu} \left(\inf_{k \geq 0} f(x_k) - \delta - f^* \right) > \epsilon. \quad (3.4)$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\text{lev}}$] the current best function value $\min_{0 \leq j \leq k} f(x_j)$ decreases by at least δ , so in view of Eq. (3.4) the target level can be attained only a finite number of times. Therefore, according to Eq. (2.7), there is a nonnegative integer k_0 such that

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

so that the target levels f_k^{lev} satisfy

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

By choosing a larger k_0 if necessary, from the preceding relation and Eq. (3.4) it can be seen that for some positive scalar ν we have

$$\frac{\mu - R}{\mu} (f_k^{\text{lev}} - f^*) \geq \epsilon_k + \nu, \quad \forall k \geq k_0. \quad (3.5)$$

Next, by using Lemma 3.1 and the definition of α_k , we obtain for all k

$$\begin{aligned} (\text{dist}(x_{k+1}, X^*))^2 &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \frac{\mu - R}{\mu} (f(x_k) - f^*) + 2\alpha_k \epsilon_k + \alpha_k \gamma_k (f(x_k) - f_k^{\text{lev}}) \\ &= (\text{dist}(x_k, X^*))^2 - \alpha_k \left(2 \frac{\mu - R}{\mu} - \gamma_k \right) (f(x_k) - f_k^{\text{lev}}) \\ &\quad - 2\alpha_k \left(\frac{\mu - R}{\mu} (f_k^{\text{lev}} - f^*) - \epsilon_k \right) \\ &\leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \left(\frac{\mu - R}{\mu} (f_k^{\text{lev}} - f^*) - \epsilon_k \right), \end{aligned} \quad (3.6)$$

where in the last inequality above we use the facts $f(x_k) - f_k^{\text{lev}} \geq \delta_k > 0$ and $2(\mu - R)/\mu - \gamma_k \geq 0$ for all k . By substituting Eq. (3.5) in the preceding inequality, we have for all $k \geq k_0$

$$(\text{dist}(x_{k+1}, X^*))^2 \leq (\text{dist}(x_k, X^*))^2 - 2\alpha_k \nu \leq \dots \leq (\text{dist}(x_{k_0}, X^*))^2 - 2\nu \sum_{j=k_0}^k \alpha_j, \quad (3.7)$$

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 $f(x_k) - f_k^{\text{lev}} \geq \delta$ for all k , from the definition of α_k we obtain

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

which when substituted in Eq. (3.7) yields a contradiction for sufficiently large k . **Q.E.D.**

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

Proposition 3.5: Let Assumptions 3.1 and 3.3 hold. Furthermore, let the parameters γ_k in the path-based procedure be such that

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

Then, for a sequence $\{x_k\}$ generated by the method and the path-based procedure, we have

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

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

$$\inf_{k \geq 0} f(x_k) > f^* + \frac{\mu \epsilon}{\mu - R}. \quad (3.8)$$

If 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 α_k we obtain

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

contradicting the fact $\alpha_k \|\tilde{g}_k\| \rightarrow 0$. Hence $l \rightarrow \infty$. If δ_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 $\min_{0 \leq j \leq k} f(x_j)$ is decreased by at least $\delta_{\bar{l}}/2$, so in view of Eq. (3.8) there can be only a finite number of sufficient descents, which is a contradiction. Hence δ_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 ν be a positive scalar such that [cf. Eq. (3.8)]

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

Then by using the fact $f_k^{\text{lev}} = \min_{0 \leq j \leq k(l)} -\delta_l$, we obtain

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

Similar to the proof of Prop. 3.4, it can be seen that for all k we have [cf. Eq. (3.6)]

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

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

$$\sum_{k=0}^{\infty} \alpha_k < \infty \tag{3.9}$$

and the sequence $\{\text{dist}(x_k, X^*)\}$ is bounded. By Assumption 3.3, it follows that 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 .

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

Therefore

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j \|\tilde{g}_j\| > B, \quad \forall l \in L, \tag{3.10}$$

and by using the fact $\|\tilde{g}_k\| \leq C + R$ for all k , we obtain

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

Because the cardinality of L is infinite, we have

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

contradicting the relation $\sum_{k=0}^{\infty} \alpha_k < \infty$ [cf. Eq. (3.9)]. **Q.E.D.**

We now discuss how the noise r_k and the ϵ_k errors affect the error estimate results established in this section. 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 ϵ -subgradients are used (i.e., $\epsilon = 0$) and the low level noise is persistent, the error in the estimates of Props. 3.2, 3.3, and 3.5 vanishes and the convergence to the exact value f^* is obtained. By contrast, exact convergence cannot be guaranteed in the results of Section 2. In particular, the error estimates of Props. 2.2, 2.3, and 2.5 are persistent (do not diminish to zero) even when $\epsilon = 0$, and thus only convergence to an approximation of the value f^* can be guaranteed (compare Props. 2.2, 2.3, and 2.5 with Props. 3.2, 3.3, and 3.5, respectively).

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, 3.3, and 3.5 does not vanish. In particular, the resulting error is proportional to the limiting error ϵ associated with ϵ_k -subgradients used in the method. This demonstrates the different nature of the noise r_k and the errors associated with using ϵ_k -subgradients.

We illustrate the persistency of the effects of ϵ -errors in the following example, which also shows that our estimate of Prop. 3.2 is tight.³

Example 4.1:

Consider the problem of minimizing $f(x) = |x|$ over $x \in \mathfrak{R}$. The set of minima is $X^* = \{0\}$ and the sharp minima parameter μ is equal to 1. For $\epsilon > 0$, it can be seen that

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

³ 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>.

Let the ϵ -subgradient noise bound be R with $0 \leq R < 1$. Consider the ϵ -subdifferential $\partial_\epsilon f(x_0)$ at a point $x_0 = \frac{\epsilon}{1-R}$. We have

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

Thus, $g = R$ is an ϵ -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 ϵ is persistent, and also makes the estimate of Prop. 3.2 sharp (for $\mu = 1$).

4. IMPLICATIONS FOR INCREMENTAL ϵ_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 $f_i : \mathfrak{R}^n \rightarrow \mathfrak{R}$ 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 [Kib79], Nedić and Bertsekas [NeB00], [NeB01], Nedić, Bertsekas, and Borkar [NBB01] (see also Nedić [Ned02]), Ben-Tal, Margalit, and Nemirovski [BMN01], and Kiwiel [Kiw04]. The presence of noise in incremental subgradient methods was addressed by Solodov and Zavriev in [SoZ98] for a compact constraint set X and the diminishing stepsize rule. See [BNO03] for an extensive reference on incremental subgradient methods.

Convergence Results for a Compact Constraint Set

Here, we show that the results of Section 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 $i = 1, \dots, m$,

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

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 ϵ -subdifferential sets are nested as ϵ 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 [Dev85], Corollary on pg. 77), implying that there exist a constant C_i such that for all $x \in X$,

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

Note that, since $\partial_g(x) \subseteq \partial_\epsilon g(x)$ for any convex function g and any x , it follows that the subgradients of the functions f_i are also bounded by the same constants respectively, i.e., for all $i = 1, \dots, m$ and $x \in X$,

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

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 nonincremental 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

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

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 ,

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

where the last inequality follows from the fact

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

[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

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

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 , we can strengthen the above inequality as follows

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

After some calculation, it can be seen that

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

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

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

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

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

Q.E.D.

We consider the incremental method using the constant, the diminishing, and the modified dynamic stepsize rules. The modification of the dynamic stepsize rules consists of replacing $\|\tilde{g}_k\|^2$ by $(\tilde{C} + \tilde{R})^2 - \tilde{S}$ in Eqs. (1.5) and (1.6), and at Step 5 of the path-based procedure, the parameter σ_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 α_k .

Under Assumptions 2.1 and 4.1, we can show that the results of Props. 2.1–2.5 apply to a sequence $\{x_k\}$ generated by the incremental method, where in the estimates of Section 2 we replace $(C + R)^2$ by $(\tilde{C} + \tilde{R})^2 - \tilde{S}$ and $\epsilon = \sum_{i=1}^m \epsilon_i$ (with ϵ_i as defined in Assumption 4.1). This can be seen by using Lemma 4.1 in place of Lemma 2.1.

Convergence Results for f with Sharp Set of Minima

In this section, we show that the results of Section 3 also hold for an incremental ϵ_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 ϵ -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

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

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

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

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. **Q.E.D.**

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, under Assumptions 4.1 and 4.2, we can show that the results of Props. 3.1–3.5 apply to a sequence $\{x_k\}$ generated by the incremental method. In this case, the results of Section 3 hold [with $(\tilde{C} + \tilde{R})^2 - \tilde{S}$ instead of $(C + R)^2$ in the case of Prop. 3.1], and $\epsilon = \sum_{i=1}^m \epsilon_i$ (with ϵ_i as in Assumption 4.1). This can be seen by using Lemma 4.2 in place of Lemma 3.1 and a line of analysis identical to that of Section 3.

5. REFERENCES

- [BMN01] Ben-Tal A., Margalit T., and Nemirovski A., “The Ordered Subsets Mirror Descent Optimization Method and its Use for the Positron Emission Tomography Reconstruction,” *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, Eds. D. Butnariu, Y. Censor and S. Reich, Studies in Comput. Math., Elsevier, 2001.
- [BNO03] Bertsekas D. P., Nedić A., and Ozdaglar A. E., *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.
- [Brä93] Brännlund U., “On Relaxation Methods for Nonsmooth Convex Optimization,” *Doctoral Thesis*, Royal Institute of Technology, Stockholm, Sweden, 1993.
- [BuF93] Burke J. V. and Ferris M. C., “Weak sharp minima in mathematical programming,” *SIAM J. on Control and Optim.*, Vol. 31, No. 5, 1993, pp. 1340–1359.

- [DeV85] Dem'yanov V. F. and Vasil'ev L. V., *Nondifferentiable Optimization*, Optimization Software Inc., New York, 1985.
- [Erm69] Ermoliev Yu. M., "On the Stochastic Quasi-Gradient Method and Stochastic Quasi-Feyer Sequences," *Kibernetika*, No. 2, 1969, pp. 73–83.
- [Erm76] Ermoliev Yu. M., *Stochastic Programming Methods*, Nauka, Moscow, 1976.
- [Erm83] Ermoliev Yu. M., "Stochastic Quasigradient Methods and Their Application to System Optimization," *Stochastics*, Vol. 9, 1983, pp. 1–36.
- [Erm88] Ermoliev Yu. M., "Stochastic Quasigradient Methods," in *Numerical Techniques for Stochastic Optimization*, Eds., Yu. M. Ermoliev and R. J-B. Wets, IIASA, Springer-Verlag, 1988, pp. 141–185.
- [GoK99] Goffin J. L. and Kiwiel K., "Convergence of a Simple Subgradient Level Method," *Math. Programming*, Vol. 85, 1999, pp. 207–211.
- [Kib79] Kibardin V. M., "Decomposition into Functions in the Minimization Problem," *Automation and Remote Control*, Vol. 40, 1980, pp. 1311–1323.
- [Kiw04] Kiwiel K. C., "Convergence of Approximate and Incremental Subgradient Methods for Convex Optimization," *SIAM J. on Optimization*, Vol. 14, No. 3, 2004, pp. 807–840.
- [NBB01] Nedić A., Bertsekas D. P., and Borkar V., "Distributed Asynchronous Incremental Subgradient Methods," *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, Eds. D. Butnariu, Y. Censor and S. Reich, *Studies in Comput. Math.*, Elsevier, 2001.
- [NeB00] Nedić A. and Bertsekas D. P., "Convergence Rate of Incremental Subgradient Algorithm," *Stochastic Optimization: Algorithms and Applications*, Eds., S. Uryasev and P. M. Pardalos, Kluwer Academic Publishers, 2000, pp. 263–304.
- [NeB01] Nedić A. and Bertsekas D. P., "Incremental Subgradient Methods for Nondifferentiable Optimization," *SIAM J. on Optimization*, Vol. 12, 2001, pp. 109–138.
- [Ned02] Nedić A., "Subgradient Methods for Convex Optimization," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.
- [Nur74] "Minimization of Nondifferentiable Functions in Presence of Noise," *Kibernetika*, Vol. 10, No. 4, 1974, pp. 59–61.
- [Pol87] Polyak B. T., "Nonlinear Programming Methods in the Presence of Noise," *Math. Programming*, Vol. 14, 1978, pp. 87–97.

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

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