SIAM J. OPTIM.
Vol. 14, No. 3, pp. 807–840

CONVERGENCE OF APPROXIMATE AND INCREMENTAL SUBGRADIENT METHODS FOR CONVEX OPTIMIZATION*

KRZYSZTOF C. KIWIEL†

Abstract. We present a unified convergence framework for approximate subgradient methods that covers various stepsize rules (including both diminishing and nonvanishing stepsizes), convergence in objective values, and convergence to a neighborhood of the optimal set. We discuss ways of ensuring the boundedness of the iterates and give efficiency estimates. Our results are extended to incremental subgradient methods for minimizing a sum of convex functions, which have recently been shown to be promising for various large-scale problems, including those arising from Lagrangian relaxation.

Key words. nondifferentiable optimization, convex programming, subgradient optimization, approximate subgradients, efficiency

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/S1052623400376366

1. Introduction.

We are interested in the convex constrained minimization problem

$$f_* := \inf \{ f(x) : x \in S \} \quad \text{with} \quad f := \sum_{i=1}^{m} f_i, \tag{1.1}$$

where $S \neq \emptyset$ is a closed convex set in the Euclidean space $\mathbb{R}^n$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, and each $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed proper convex function finite on S. Let $S_* := \operatorname{Arg\,min}_S f$ denote the *optimal set* of problem (1.1) and $f_S := f + \mathrm{I}_S$ its *extended objective*, where I_S is the *indicator function* of S ($\mathrm{I}_S(x) = 0$ if $x \in S$, ∞ if $x \notin S$). Then $f_* = \inf f_S$ and $S_* = \operatorname{Arg\,min} f_S$; note that f_S is a closed proper convex function.

The *approximate subgradient projection method* generates a sequence $\{x^k\}_{k=1}^{\infty} \subset S$ via

$$x^{k+1} := P_S(x^k - \nu_k g^k), \quad g^k \in \partial_{\epsilon_k} f_S(x^k), \quad k = 1, 2, \dots, \quad x^1 \in S, \tag{1.2}$$

where $P_S x := \arg\min_S |x - \cdot|$ is the *projector* on S, $\nu_k > 0$ is a *stepsize*, and $\epsilon_k \geq 0$ is an error tolerance of an approximate subgradient g^k that belongs to the ϵ_k-subdifferential of f_S at x^k:

$$\partial_{\epsilon_k} f_S(x^k) := \left\{ g : f_S(x) \geq f_S(x^k) + \left\langle g, x - x^k \right\rangle - \epsilon_k \quad \forall x \right\}. \tag{1.3}$$

This method, introduced by Shor [Sho62] and first analyzed in [Erm66, Pol67] has extensive literature; see, e.g., the books [Ber99, BSS93, DeV81, Min86, Nes89, Pol83, Sho79] (and, e.g., [Erm76, MGN87, Nur79] for extensions to stochastic and nonconvex problems). However, most authors tailor their analyses to particular stepsizes, such as $\nu_k := \lambda_k |g^k|^{-1}$ with $\sum_k \lambda_k = \infty$.

*Received by the editors August 4, 2000; accepted for publication (in revised form) July 2, 2002; published electronically March 5, 2004. This research was supported by the State Committee for Scientific Research under grant 8T11A00622.
http://www.siam.org/journals/siopt/14-3/37636.html

†Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

This paper presents a unified convergence framework for the method (1.2) that covers various stepsize rules (including both diminishing and nonvanishing stepsizes), convergence in the objective values $f(x^k)$, and convergence of $\{x^k\}$ to the optimal set S_* or its neighborhood for nonvanishing stepsizes. We discuss ways of ensuring boundedness of the iterates and give efficiency estimates. Our results subsume those in the literature.

Our analysis extends to the *incremental subgradient projection method* given by

$$x_1^k := x^k, \quad x_{i+1}^k := P_S(x_i^k - \nu_k g_i^k), \quad g_i^k \in \partial_{\epsilon_i^k} f_i^S(x_i^k), \quad i = 1\colon m, \tag{1.4a}$$

$$x^{k+1} := x_{m+1}^k, \tag{1.4b}$$

where $f_i^S := f_i + \mathrm{I}_S$. In other words, subgradient steps are taken for successive objectives f_i of (1.1), hoping that one iteration with m steps should be almost as effective as m ordinary iterations (1.2), although it is much cheaper. This hope is supported by the recent analysis and numerical results of [BTMN01, NeB01], where this version is shown to be promising for certain large-scale problems, including those arising from Lagrangian relaxation. The incremental version stems from [Kib79], but for differentiable problems it is related to backpropagation methods in neural networks; see, e.g., [Ber97, BeT00, Gai94, Gri94, Luo91, LuT94, MaS94].

The paper is organized as follows. In section 2 we recall some elementary results on ergodic convergence and coercivity. General convergence results are given in section 3, and the cases where f_S is coercive or $\{x^k\}$ is bounded are studied in sections 4 and 5. In section 6 we discuss techniques that ensure boundedness of $\{x^k\}$, whereas in section 7 we analyze stepsize rules that do not need such techniques. (Unfortunately, they do not extend to the incremental case.) Efficiency estimates for various stepsizes are given in section 8. Finally, section 9 extends the preceding convergence and efficiency results to the incremental case.

Our notation is fairly standard. $B_\rho := \{x : |x| \le \rho\}$ is the ball with center 0, and radius ρ. $d_C(\cdot) := \inf_{y \in C} |\cdot - y|$ is the distance function of a set $C \subset \mathbb{R}^n$.

2. Technical preliminaries.

We present the following three lemmas in order to make the paper more self-contained.

LEMMA 2.1. *Suppose $\nu_k > 0$ and $\nu_{\mathrm{sum}}^k := \sum_{j=1}^k \nu_j \to \infty$ as $k \to \infty$. Given a scalar sequence $\{a_k\}$, let $\bar a_k := \sum_{j=1}^k \nu_j a_j / \nu_{\mathrm{sum}}^k$ for all k. Then $\underline{\lim}_{k\to\infty} a_k \le \underline{\lim}_{k\to\infty} \bar a_k \le \overline{\lim}_{k\to\infty} \bar a_k \le \overline{\lim}_{k\to\infty} a_k$. In particular, if $\lim_{k\to\infty} a_k$ exists, then $\lim_{k\to\infty} \bar a_k = \lim_{k\to\infty} a_k$.*

Proof. To show that $a := \underline{\lim}_k a_k \le \underline{\lim}_k \bar a_k$, suppose $a > -\infty$. For any $\epsilon > 0$, pick $\bar\jmath$ such that $a_j \ge a - \epsilon$ for all $j \ge \bar\jmath$ and $\sum_{j=1}^{\bar\jmath} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k \ge -\epsilon$ for all $k \ge \bar\jmath$; then

$$\bar a_k - a = \sum_{j=1}^{\bar\jmath} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k + \sum_{j=\bar\jmath+1}^{k} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k \ge -\epsilon - \epsilon \sum_{j=\bar\jmath+1}^{k} \nu_j/\nu_{\mathrm{sum}}^k \ge -2\epsilon$$

for all $k \ge \bar\jmath$. Applying this to $b_k := -a_k$, $\bar b_k := -\bar a_k$ gives $-\overline{\lim}_k a_k \le -\overline{\lim}_k \bar a_k$. ◻

LEMMA 2.2 (Silverman–Toeplitz's theorem [DuS88, p. 75]). *Let $a_{kj} \in \mathbb{R}_+$, $j = 1\colon k$, $k = 1, 2, \ldots$, be such that $\sum_{j=1}^k a_{kj} = 1$ for all k, $\lim_{k\to\infty} a_{kj} = 0$ for all*

j (e.g., $a_{kj} = \nu_j/\nu^k_{\rm sum}$ as in Lemma 2.1). If $\{u^j\} \subset \mathbb{R}^n$ is a sequence such that $\lim_{j\to\infty} u^j = u$, then $\lim_{k\to\infty}\sum_{j=1}^k a_{kj}u^j = u$.

LEMMA 2.3. *Let $\{a_k\}$, $\{b_k\}$, and $\{c_k\}$ be sequences in $\mathbb{R}_+$ such that $a_{k+1} \le a_k(1+b_k)+c_k$ for $k = 1, 2, \ldots$, $\sum_{k=1}^\infty b_k < \infty$, $\sum_{k=1}^\infty c_k < \infty$. Then $\{a_k\}$ converges to some $a_\infty < \infty$.*

Proof. See, e.g., [Pol83, Lem. 2.2.2], due to [Gla65]. □

Denote the *trench* (sublevel set) of the extended objective f_S at any level $\alpha \in \mathbb{R}$ by

$$T_\alpha := \{\, x : f_S(x) \le \alpha \,\}. \tag{2.1}$$

Recalling that f_S is closed and convex, note that the following are equivalent: (i) f_S is *coercive*, i.e., $\lim_{|x|\to\infty} f_S(x) = \infty$; (ii) f_S is *level-bounded*; i.e., T_α is bounded for all $\alpha \in \mathbb{R}$; (iii) the optimal set $S_* = \operatorname{Arg\,min} f_S$ is nonempty and bounded [Roc70, Thm. 27.2].

We shall need some elementary properties of the trenches of f_S and their neighborhoods.

LEMMA 2.4. *Suppose that f_S is coercive and its trench T_β is nonempty for some $\beta \in \mathbb{R}$.*

(i) *For each level $\alpha \ge \beta$, let*

$$\rho(\alpha) := \max_{x\in T_\alpha} d_{T_\beta}(x) = \min\{\rho \ge 0 : T_\alpha \subset T_\beta + B_\rho\} \quad \text{and} \quad T^\alpha_\beta := T_\beta + B_{\rho(\alpha)}; \tag{2.2}$$

thus $\rho(\alpha)$ is the distance between T_α and T_β, whereas T^α_β is the smallest neighborhood of T_β containing T_α, so that $T_\beta \subset T^\alpha_\beta \subset T_\beta + B_\rho$ whenever $\rho \ge \rho(\alpha)$. Then $\lim_{\alpha\downarrow\beta}\rho(\alpha) = 0$.

(ii) *If f_S is also continuous on its domain S (i.e., f is continuous on S), then for every level $\bar\alpha > \beta$ there exists a radius $\bar\rho > 0$ such that $S \cap (T_\beta + B_{\bar\rho}) \subset T_{\bar\alpha}$.*

Proof. (i) Since f_S is closed and coercive, both T_β and T_α are compact, and $\rho(\alpha)$ is well defined by (2.2) (d_{T_β} is continuous) and nondecreasing (so is T_α by (2.1)). To show that $\lim_{\alpha\downarrow\beta}\rho(\alpha) = 0$ by contradiction, suppose there are sequences $\alpha_i \downarrow \beta$ and $y^i \in T_{\alpha_i}$ such that $d_{T_\beta}(y^i) \ge \rho > 0$. Since $T_{\beta+1}$ is bounded, we may assume without loss of generality that $y^i \to y^\infty$. Then $d_{T_\beta}(y^\infty) \ge \rho$, since d_{T_β} is continuous. However, $f_S(y^i) \le \alpha_i$ gives in the limit $f_S(y^\infty) \le \beta$ (f_S is closed) and hence $y^\infty \in T_\beta$, contradicting $d_{T_\beta}(y^\infty) \ge \rho$.

(ii) Otherwise there are $\rho_i \downarrow 0$, $y^i \in S \cap (T_\beta + B_{\rho_i}) \setminus T_{\bar\alpha}$, $z^i \in T_\beta$ such that $|y^i - z^i| \le \rho_i$. Since T_β is compact, we may assume without loss of generality that $z^i \to z^\infty \in T_\beta$. However, then $y^i \to z^\infty$ (since $|y^i - z^i| \to 0$) with $f_S(y^i) \ge \bar\alpha$ ($y^i \notin T_{\bar\alpha}$) and the continuity of f_S on S imply $f_S(z^\infty) \ge \bar\alpha$, which contradicts $z^\infty \in T_\beta$. □

LEMMA 2.5. *Suppose that f_S is coercive, $\sigma \in [0,\infty)$, and $\alpha \in \mathbb{R}$. Then the set*

$$T_{\alpha,\sigma} := \left\{ x : \sum_{i=1}^m f_i^S(x_i) \le \alpha \text{ for some } x_i \in x + B_\sigma \right\} \tag{2.3}$$

is bounded.

Proof. For each i, let $\hat f_i(x) := \inf_{y\in x+B_\sigma} f_i^S(y) = \inf_y\{f_i^S(y) + \mathrm{I}_{B_\sigma}(x-y)\}$ for all x. Since each f_i^S is closed proper convex, so is $\hat f_i$, and they have the same recession function (cf. [Roc70, Cors. 9.2.1 and 9.2.2]). Hence (cf. [Roc70, Thm. 9.3]) $f_S = \sum_i f_i^S$ and $\hat f := \sum_i \hat f_i$ have a common recession function and a common recession cone. This cone is null because f_S is coercive, so $\hat{f}$ is coercive (cf. [Roc70, Thms. 8.4 and 8.7]); hence its level set $\{x : \hat{f}(x) \le \alpha\}$ is bounded. This set coincides with $T_{\alpha,\sigma}$, since $\hat{f}_i(x) \le \alpha_i$ iff $f_i^S(x_i) \le \alpha_i$ for some $x_i \in x + B_\sigma$, because f_i^S is closed and the ball $x + B_\sigma$ is compact. ∎

3. General convergence results. Throughout this section, and in the following sections until section 9, $\{x^k\}$, $\{\nu_k\}$, $\{\epsilon_k\}$, and $\{g^k\}$ denote the sequences involved in the (ordinary) subgradient iteration (1.2).

3.1. Basic estimates. Our convergence analysis hinges on the following three simple estimates.

LEMMA 3.1. *For each x and $k \ge 1$, we have*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f_S(x) - \epsilon_k - \tfrac{1}{2}|g^k|^2\nu_k \right], \tag{3.1}$$

$$\frac{\sum_{j=1}^k \nu_j f(x^j)}{\sum_{j=1}^k \nu_j} - f_S(x) \le \frac{\frac{1}{2}|x^1 - x|^2 + \sum_{j=1}^k \frac{1}{2}\nu_j^2 |g^j|^2 + \sum_{j=1}^k \nu_j \epsilon_j}{\sum_{j=1}^k \nu_j}, \tag{3.2}$$

$$|x^{k+1} - x^k| \le \nu_k |g^k|. \tag{3.3}$$

Proof. Let $x \in S$, $r_k := |x^k - x|$. Using the nonexpansiveness of P_S and (1.2)–(1.3) gives

$$\begin{aligned} r_{k+1}^2 &\le |x^k - \nu_k g^k - x|^2 = r_k^2 - 2\nu_k \left\langle g^k, x^k - x \right\rangle + \nu_k^2 |g^k|^2 \\ &\le r_k^2 + 2\nu_k \left[f_S(x) - f(x^k) + \epsilon_k \right] + \nu_k^2 |g^k|^2, \end{aligned} \tag{3.4}$$

and hence (3.1). Summing up (3.1) yields (3.2). For $f_S(x) = \infty$, (3.1)–(3.2) are trivial. Finally, (3.3) follows from the nonexpansiveness of P_S and the fact that $x^k \in S$ in (1.2). ∎

Denoting the quantities involved in the basic estimate (3.1) by

$$\gamma_k := \tfrac{1}{2}|g^k|^2 \nu_k \quad \text{and} \quad \delta_k := \gamma_k + \epsilon_k, \tag{3.5}$$

we have

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f_S(x) - \delta_k \right] \quad \forall x. \tag{3.6}$$

Thus x^{k+1} becomes closer than x^k to points x such that $f(x^k) > f_S(x) + \delta_k$, and it is easy to see that the standard stepsize condition $\sum_k \nu_k = \infty$ yields $\underline{\lim}_k f(x^k) \le f_S(x) + \delta$ for all x and hence $\underline{\lim}_k f(x^k) \le f_* + \delta$ for $\delta := \overline{\lim}_k \delta_k$. (Of course, additional assumptions are needed to ensure $\delta < \infty$.) In fact, stronger results are derived in the next subsection by employing averages of $\{x^k\}$ and $\{\delta_k\}$ weighted by the stepsizes $\{\nu_k\}$.

3.2. Cesáro averages and ergodic convergence. Employing, as usual, an unbounded *summary stepsize*

$$\nu_{\mathrm{sum}}^k := \sum_{j=1}^k \nu_j \to \infty \quad \text{as } k \to \infty, \tag{3.7}$$

we shall study the *Cesáro averages* of the sequences $\{x^k\}$ and $\{f(x^k)\}$ defined by

$$\bar{x}^k := \sum_{j=1}^{k} \nu_j x^j/\nu_{\mathrm{sum}}^k \quad \text{and} \quad \bar{f}_k := \sum_{j=1}^{k} \nu_j f(x^j)/\nu_{\mathrm{sum}}^k. \tag{3.8}$$

Note that, since $\nu_k > 0$ and $x^k \in S$, for all k, the convexity of f, S, and $|\cdot|$ yields

$$f(\bar{x}^k) \le \bar{f}_k, \quad \bar{x}^k \in S, \quad \text{and} \quad |\bar{x}^k| \le \max\{\,|x^j| : j = 1\colon k\,\}. \tag{3.9}$$

Using the Cesáro averages of the sequences $\{\gamma_k\}$, $\{\epsilon_k\}$, and $\{\delta_k\}$ (cf. (3.5)),

$$\bar{\gamma}_k := \sum_{j=1}^{k} \nu_j \gamma_j/\nu_{\mathrm{sum}}^k, \quad \bar{\epsilon}_k := \sum_{j=1}^{k} \nu_j \epsilon_j/\nu_{\mathrm{sum}}^k, \quad \text{and} \quad \bar{\delta}_k := \sum_{j=1}^{k} \nu_j \delta_j/\nu_{\mathrm{sum}}^k = \bar{\gamma}_k + \bar{\epsilon}_k, \tag{3.10}$$

we may rewrite the estimate (3.2) in the Cesáro average form

$$\bar{f}_k - f_S(x) \le \tfrac{1}{2}|x^1 - x|^2/\nu_{\mathrm{sum}}^k + \bar{\delta}_k \quad \forall x. \tag{3.11}$$

It is convenient to employ the shorthand notation

$$\bar{\gamma}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\gamma}_k, \quad \bar{\epsilon}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\epsilon}_k, \quad \bar{\delta}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\delta}_k, \quad \text{and} \quad \bar{\delta}_{\inf} := \underline{\lim_{k\to\infty}}\, \bar{\delta}_k. \tag{3.12}$$

For each $\delta \ge 0$, denote the set of *δ-optimal points* of problem (1.1) by

$$S_\delta := \{\, x : f_S(x) \le f_* + \delta \,\}. \tag{3.13}$$

We now show that the algorithm attempts asymptotically to find points in the set $S_{\bar{\delta}_{\sup}}$.

THEOREM 3.2. *Assuming $\sum_{k=1}^\infty \nu_k = \infty$, define $\bar{\delta}_{\sup}$ and $\bar{\delta}_{\inf}$ by* (3.12) *and* (3.10). *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(\bar{x}^k) \le \underline{\lim}_{k\to\infty} \bar{f}_k \le f_* + \bar{\delta}_{\inf}$ *and* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar{\delta}_{\inf}$.

(ii) $\overline{\lim}_{k\to\infty} f(\bar{x}^k) \le \overline{\lim}_{k\to\infty} \bar{f}_k \le f_* + \bar{\delta}_{\sup}$ *and* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar{\delta}_{\sup}$.

(iii) *If $\bar{\delta}_{\sup} = 0$, then $f(\bar{x}^k)$, $\bar{f}_k$, and $\inf_{l\ge k} f(x^l)$ converge to f_* as $k \to \infty$.*

(iv) *All the cluster points of $\{\bar{x}^k\}$ (if any) lie in the $\bar{\delta}_{\sup}$-optimal set $S_{\bar{\delta}_{\sup}}$.*

(v) *If $S_* = \emptyset$ and $\bar{\delta}_{\sup} = 0$, then $|\bar{x}^k| \to \infty$ and $\overline{\lim}_{k\to\infty} |x^k| = \infty$.*

(vi) *$\bar{\delta}_{\sup} \le \bar{\gamma}_{\sup} + \bar{\epsilon}_{\sup}$, $\bar{\delta}_{\sup} \le \overline{\lim}_{k\to\infty} \delta_k$, $\bar{\gamma}_{\sup} \le \overline{\lim}_{k\to\infty} \gamma_k$, and $\bar{\epsilon}_{\sup} \le \overline{\lim}_{k\to\infty} \epsilon_k$. In particular, $\bar{\gamma}_{\sup} = 0$ if $\lim_{k\to\infty} \nu_k |g^k|^2 = 0$ (e.g., $\lim_{k\to\infty} \nu_k = 0$ and $\sup_k |g^k| < \infty$). If $\nu := \overline{\lim}_{k\to\infty} \nu_k$ and $C := \overline{\lim}_{k\to\infty} |g^k|$ are finite, then $\bar{\gamma}_{\sup} \le \frac{1}{2} C^2 \nu < \infty$.*

Proof. (i) Since by assumption $\nu_{\mathrm{sum}}^k \to \infty$, taking lower limits in (3.11) gives $\underline{\lim}_k \bar{f}_k \le f_S(x) + \bar{\delta}_{\inf}$ for each x, so $f_* := \inf f_S$ yields $\underline{\lim}_k \bar{f}_k \le f_* + \bar{\delta}_{\inf}$. The conclusion follows from the facts that $f(\bar{x}^k) \le \bar{f}_k$ for all k (cf. (3.9)) and $\underline{\lim}_k f(x^k) \le \underline{\lim}_k \bar{f}_k$ (cf. Lemma 2.1).

(ii) Argue as for (i), replacing lower limits by upper limits.

(iii) This follows from (ii), since (cf. (3.8)–(3.9)) $f(x^k), f(\bar{x}^k), \bar{f}_k \ge f_*$.

(iv) If $\{\bar{x}^k\}$ has a cluster point $\bar{x}^\infty$, then $f_S(\bar{x}^\infty) \le f_* + \bar{\delta}_{\sup}$ by (ii), since f_S is closed.

(v) If $|\bar{x}^k| \not\to \infty$, then $\{\bar{x}^k\}$ has a cluster point $\bar{x}^\infty$ in $S_0 = S_*$ by (iv), i.e., $S_* \ne \emptyset$. Hence if $S_* = \emptyset$, then $|\bar{x}^k| \to \infty$, with $|\bar{x}^k| \le \max_{j=1}^k |x^j|$ by (3.9).

(vi) This follows from (3.12), (3.10), (3.5), (3.7), and Lemma 2.1. $\square$

Remark 3.3.

(i) Theorem 3.2 implies additional results for the *record* points

$$x_{\rm rec}^k \in \operatorname{Arg}\min_{\{x^j\}_{j=1}^k} f(x^j) \subset S \quad \text{with} \quad f(x_{\rm rec}^k) = \min_{j=1:\,k} f(x^j) \le \bar f_k, \tag{3.14}$$

where the inequality stems from (3.7)–(3.8). Specifically, $x_{\rm rec}^k$ may replace $\bar x^k$ throughout, also with $\bar\delta_{\rm sup}$ replaced by $\bar\delta_{\rm inf}$ in parts (iii)–(v). However, $\bar x^k = (\nu_k x^k + \nu_{\rm sum}^{k-1}\bar x^{k-1})/\nu_{\rm sum}^k$ may be updated at negligible cost *without* evaluating f, in contrast with $x_{\rm rec}^k$.

(ii) Theorem 3.2 handles both diminishing stepsizes ($\nu = 0$ in (vi)) and nonvanishing ones ($\nu > 0$), for which $\nu_k|g^k|^2 \to 0$ is unlikely in the nonsmooth case.

(iii) The second part of Theorem 3.2(ii) subsumes [Ber99, Ex. 6.3.13(a)] (where $\epsilon_k \to \bar\epsilon$ and $\nu_k|g^k|^2 \to 0$ so that $\bar\delta_{\rm sup} = \bar\epsilon$), which in turn generalizes [CoL93, Prop. 1.2] (where $\bar\epsilon = 0$); its first part subsumes [MiU82, Thm. 1] (where $\nu_k \to 0$, $\sup_k |g^k| < \infty$, and $\epsilon_k \equiv 0$).

3.3. Full convergence. To ensure convergence of $\{x^k\}$, we need stronger assumptions (relative to Theorem 3.2).

THEOREM 3.4. *Suppose* $\sum_{k=1}^\infty \nu_k = \infty$, $\sum_{k=1}^\infty \nu_k\delta_k < \infty$ (*cf.* (3.5)). *Then the conclusions of Theorem* 3.2(i–v) *hold with* $\bar\delta_{\rm sup} = 0$*, and the following statements are equivalent*:

(i) *The optimal set* S_* *is nonempty.*

(ii) $\{x^k\}$ *is bounded* (*where* "(i) $\Rightarrow$ (ii)" *does not require* $\sum_k \nu_k = \infty$).

(iii) $\{x^k\}$ *converges to some* $x^\infty \in S_*$.

Finally, if $\{x^k\}$ *converges to a point* x^∞*, then* $\{\bar x^k\}$ *converges to the same point.*

Proof. By (3.7), (3.10), and (3.12), $\sum_k \nu_k\delta_k < \infty$ yields $\bar\delta_{\rm sup} = 0$ for Theorem 3.2.

"(i) $\Rightarrow$ (ii)": Let $x \in S_*$. Then $f_S(x) \le f(x^k)$, so the basic estimate (3.6) yields

$$|x^{k+1} - x|^2 \le |x^k - x|^2 + 2\nu_k\delta_k \quad \forall k.$$

Hence Lemma 2.3 with $b_k := 0$ and $c_k := 2\nu_k\delta_k$ shows that $a_k := |x^k - x|$ converges. Thus $\{x^k\}$ is bounded. "(i) $\Leftarrow$ (ii)": If $\{x^k\}$ is bounded, then it has a cluster point $x^\infty \in S_*$, since $\underline{\lim}_k f_S(x^k) = f_*$ by Theorem 3.2(iii) and f_S is closed.

"(i) $\Rightarrow$ (iii)": As in the proof of "(i) $\Rightarrow$ (ii)", $|x^k - x|$ converges for each $x \in S_*$, and $\{x^k\}$ has a cluster point $x^\infty \in S_*$. Taking $x = x^\infty$, we get $\underline{\lim}_k |x^k - x| = 0$, and then $|x^k - x| \to 0$, i.e., $x^k \to x^\infty$. "(i) $\Leftarrow$ (iii)": The proof is trivial.

Finally, since $\nu_{\rm sum}^k \to \infty$, $x^k \to x^\infty$ yields $\bar x^k := \sum_{j=1}^k \nu_j x^j/\nu_{\rm sum}^k \to x^\infty$ (cf. Lemma 2.2). ◻

Remark 3.5.

(i) The assumption $\sum_k \nu_k\delta_k < \infty$ of Theorem 3.4 holds if $\sum_k \nu_k^2|g^k|^2 < \infty$ (e.g., $\sum_k \nu_k^2 < \infty$ and $\sup_k |g^k| < \infty$) and $\sum_k \nu_k\epsilon_k < \infty$.

(ii) For $\epsilon_k \equiv 0$, Theorem 3.4 subsumes [Ber99, Ex. 6.3.13(b)] (where the typo $\sum_k \nu_k^2 < \infty$ should be replaced by $\sum_k \nu_k^2|g^k|^2 < \infty$), [Sch83, Thm. on p. 538] (in which the claim $f(x^k) \to f_*$ is *not* proved), and [LPS96, Thm. 2.7] (where $\sum_k \nu_k^2 < \infty$, $\sup_k |g^k| < \infty$); the earliest and much cited [Pol78] result of [Lit68, Thm. 1] (claiming that $\lim_k f(x^k) = f_*$ for $\sum_k \nu_k^2 < \infty$, $\sup_k |g^k| < \infty$) has gaps in its proof, but a result similar to Theorem 3.4 follows from [ErS68] (with $\sum_k \nu_k^2 < \infty$, $\sup_k |g^k| < \infty$). For $S_* \ne \emptyset$ and $\nu_k \to 0$, Theorem 3.4 concerning Theorem 3.2(iv) recovers a part of [NeY78, Thm. (ii)]. Finally, Theorem 3.4 subsumes [LPS00, Thm. 8] (with $\sum_k \nu_k^2 < \infty$, $\sup_k |g^k| < \infty$, $\epsilon_k \to 0$).

For stepsizes such as $\nu_k := k^{-1}$, Theorem 3.4 may seem to require the boundedness of $\{g^k\}$; in fact, the norms $|g^k|$ may grow with x^k, but not too fast, as shown below.

THEOREM 3.6. *Suppose that* $\sum_{k=1}^\infty \nu_k = \infty$, $\sum_{k=1}^\infty \nu_k^2 < \infty$, $\sum_{k=1}^\infty \nu_k\epsilon_k < \infty$, *and the subgradients satisfy the linear growth condition: there exists a constant* $c < \infty$ *such that* $|g^k|^2 \le c(1+|x^k|^2)$ *for all* k. *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) = f_*$.

(ii) *If* $S_* \ne \emptyset$, *then the assumptions of Theorem* 3.4 *are satisfied with* $\sup_k |g^k| < \infty$; *in particular,* $\{x^k\}$ *and* $\{\bar{x}^k\}$ *converge to some* $x^\infty \in S_*$ *and* $\lim_{k\to\infty} f(\bar{x}^k) = f_*$.

Proof. Suppose there exist $x \in S$ and $\bar{k}$ such that $f(x^k) \ge f(x)$ for all $k \ge \bar{k}$. Employing this inequality and the linear growth condition in the basic estimate (3.1), we obtain

$$\begin{aligned} |x^{k+1}-x|^2 &\le |x^k - x|^2 + \nu_k^2 c\left(1+|x^k|^2\right) + 2\nu_k\epsilon_k - 2\nu_k\left[\, f(x^k) - f(x)\,\right] \\ &\le |x^k-x|^2 + \nu_k^2 c\left(1+2|x^k-x|^2+2|x|^2\right) + 2\nu_k\epsilon_k \\ &= |x^k-x|^2\left(1+2c\nu_k^2\right) + \left[c(1+2|x|^2)\nu_k^2 + 2\epsilon_k\nu_k\right], \end{aligned}$$

where we used the facts that $|x^k| \le |x^k-x|+|x|$ and $(a+b)^2 \le 2(a^2+b^2)$. Hence Lemma 2.3 with $b_k := 1+2c\nu_k^2$ and $c_k := c(1+2|x|^2)\nu_k^2 + 2\epsilon_k\nu_k$ shows that $a_k := |x^k - x|$ converges. Thus $\{x^k\}$ is bounded, and $\sup_k |g^k|^2 \le c(1+\sup_k |x^k|^2) < \infty$ by the linear growth condition. Then $\sum_k \nu_k^2 < \infty$ implies $\sum_k \nu_k^2 |g^k|^2 < \infty$. Thus the assumptions of Theorem 3.4 are met, and Theorem 3.2(iii) yields $\underline{\lim}_k f(x^k) = f_*$. Since $x \in S$ was arbitrary, we obtain $\underline{\lim}_k f(x^k) \le \inf f_S = f_*$, i.e., (i). For (ii), use $x \in S_*$ above and Theorem 3.4. ∎

Remark 3.7. For $S = \mathbb{R}^n$ and $\epsilon_k \equiv 0$, Theorem 3.6 recovers [PoT73, Thm. 9.1] (in the finite-dimensional deterministic setting); note that in this case $f(x^k) \to f_*$ when $x^k \to x^\infty$ by continuity of f. Again, the earliest result of [Lit68, Thm. 2] has gaps in its proof.

4. Convergence in the coercive case. We now consider the case where "everything is bounded," including the solution set S_* and the algorithmic quantities δ_k and $|x^{k+1} - x^k|$. It turns out that the asymptotic objective accuracy $\delta := \overline{\lim}_k \delta_k$ and steplength $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$ determine the neighborhood S_*^δ of S_* (cf. (4.1)) to which $\{x^k\}$ converges. The size of this neighborhood depends on the asymptotic steplength σ and on the shape of the δ-optimal set S_δ. The Cesáro averages $\{\bar{x}^k\}$ converge to the smaller set S_δ; thus averaging enhances stability.

THEOREM 4.1. *Suppose that* $\sum_{k=1}^\infty \nu_k = \infty$, $\delta := \overline{\lim}_{k\to\infty} \delta_k < \infty$, $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| < \infty$, *and* f_S *is coercive. Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} d_{S_\delta}(x^k) = 0$ *and* $\{x^k\}$ *has a cluster point in* S_δ. *Further, the assertions of Theorem* 3.2(ii)–(iii) *hold with* $\bar{\delta}_{\sup} \le \delta$.

(ii) $\lim_{k\to\infty} d_{S_*^\delta}(x^k) = 0$, *where* S_*^δ *is the neighborhood of* S_* *defined by (cf. Lemma* 2.4(i))

$$S_*^\delta := S_* + B_{\rho_\delta + \sigma} \quad \textit{with} \quad \rho_\delta := \max\left\{ d_{S_*}(x) : x \in S_\delta \right\}. \tag{4.1}$$

Thus $\{x^k\}$ *is bounded and its cluster points belong to* S_*^δ.

(iii) $\{\bar{x}^k\}$ *is bounded, its cluster points lie in* S_δ, *and* $\lim_{k\to\infty} d_{S_\delta}(\bar{x}^k) = 0$.

(iv) *In general, for* $\gamma := \overline{\lim}_{k\to\infty} \gamma_k$, $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k$, $\nu := \overline{\lim}_{k\to\infty} \nu_k$, $C := \overline{\lim}_{k\to\infty} |g^k|$, *and* $\bar{\sigma} := \overline{\lim}_{k\to\infty} \nu_k |g^k|$, *we have* $\delta \le \gamma + \epsilon$, $\gamma \le \frac{1}{2} C^2 \nu$, *and* $\sigma \le \bar{\sigma} \le \min\{C\nu, (2\gamma\nu)^{1/2}\}$. *In particular,* $\gamma = 0$ *if* $\nu = 0$ *and* $C < \infty$, *whereas* $\sigma = 0$ *if* $\bar{\sigma} = 0$ *(e.g.,* $\nu = 0$ *and* $C < \infty$, *or* $\gamma = 0$ *and* $\nu < \infty$*).*

Proof. First, recall from section 2 that the closedness and coercivity of f_S imply that the sets $S_* \subset S_\delta \subset S_* + B_{\rho_\delta} \subset S_*^\delta$ are nonempty and compact (cf. (2.2), (3.13), and (4.1)).

(i) By our assumptions and Theorem 3.2(vi), $\bar\delta_{\sup} \le \delta$. Hence Theorem 3.2(ii) gives $\underline{\lim}_k f_S(x^k) \le f_* + \delta$. Pick a subsequence $\{x^{k_j}\}$ such that $\lim_j f_S(x^{k_j}) = \underline{\lim}_k f_S(x^k)$. Since f_S is coercive, $\{x^{k_j}\}$ is bounded. Assume without loss of generality that $x^{k_j} \to x^\infty$. Then $f_S(x^\infty) \le f_* + \delta$ (f_S is closed) gives $x^\infty \in S_\delta$ (cf. (3.13)), so $d_{S_\delta}(x^{k_j}) \to d_{S_\delta}(x^\infty) = 0$ by continuity of d_{S_δ}. Thus $\underline{\lim}_k d_{S_\delta}(x^k) = 0$.

(ii) Fixing $\rho > 0$, let

$$V_{2\rho} := S_*^\delta + B_{2\rho} = \left\{ x : d_{S_*^\delta}(x) \le 2\rho \right\} \tag{4.2}$$

and

$$v_\rho := \min\left\{ f_S(x) : d_{S_\delta}(x) \ge \rho \right\} - (\, f_* + \delta\,). \tag{4.3}$$

Since f_S is closed and coercive, whereas d_{S_δ} is continuous, the minimum in (4.3) is attained at some x, and $v_\rho > 0$. (Otherwise $f_S(x) \le f_* + \delta$ would give $x \in S_\delta$ and hence $d_{S_\delta}(x) = 0$, contradicting $\rho > 0$ in (4.3).)

Since $\delta := \overline{\lim}_k \delta_k$ and $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$, there is $k_\rho < \infty$ such that

$$\delta_k \le \delta + v_\rho \quad \text{and} \quad |x^{k+1} - x^k| \le \sigma + \rho \quad \forall k \ge k_\rho. \tag{4.4}$$

Since $\underline{\lim}_k d_{S_\delta}(x^k) = 0$ by (i), there exists $k = k'_\rho \ge k_\rho$ such that $x^k \in S_\delta + B_\rho$; then $S_\delta \subset S_*^\delta$ implies $x^k \in V_{2\rho}$ (cf. (4.2)).

Assuming $x^k \in V_{2\rho}$ for some $k \ge k'_\rho$, we now show that $x^{k+1} \in V_{2\rho}$. If $d_{S_\delta}(x^k) \le \rho$, then from $S_\delta \subset S_* + B_{\rho_\delta}$, (4.1), and the second inequality of (4.4) we get

$$x^{k+1} \in (S_\delta + B_\rho) + B_{\sigma+\rho} \subset S_* + B_{\rho_\delta} + B_{\sigma+2\rho} = (S_* + B_{\rho_\delta+\sigma}) + B_{2\rho} = S_*^\delta + B_{2\rho},$$

so $x^{k+1} \in V_{2\rho}$ (cf. (4.2)). Thus suppose $d_{S_\delta}(x^k) > \rho$. Then, by (4.3),

$$f(x^k) \ge v_\rho + f_* + \delta. \tag{4.5}$$

Next, by (4.1) and (4.2),

$$V_{2\rho} = S_* + B_{\rho_\delta+\sigma+2\rho}, \tag{4.6}$$

so, since $x^k \in V_{2\rho}$, $|x^k - x| \le \rho_\delta + \sigma + 2\rho$ for $x = P_{S_*} x^k$. Using the basic estimate (3.6) with $f_S(x) = f_*$, the bound (4.5), and the first inequality of (4.4) yields

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[\, v_\rho + \delta - \delta_k \,\right] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le \rho_\delta + \sigma + 2\rho$ with $x \in S_*$, so $x^{k+1} \in V_{2\rho}$ by (4.6).

Therefore, by induction for each $k \ge k'_\rho$, $x^k \in V_{2\rho}$ and hence (cf. (4.2)) $d_{S_*^\delta}(x^k) \le 2\rho$. Since $\rho > 0$ was arbitrary, $d_{S_*^\delta}(x^k) \to 0$. Thus, since S_*^δ is bounded, so is $\{x^k\}$, and its cluster points must lie in S_*^δ because $d_{S_*^\delta}(x^k) \to 0$, $d_{S_*^\delta}$ is continuous and S_*^δ is closed.

(iii) Since $\{x^k\}$ is bounded by (ii), so is $\{\bar x^k\}$ by (3.9). Pick $\bar x^{k_j}$ such that $\lim_j d_{S_\delta}(\bar x^{k_j}) = \overline{\lim}_k d_{S_\delta}(\bar x^k)$. Extracting a subsequence if necessary, suppose $\bar x^{k_j} \to \bar x^\infty$. By Theorem 3.2(iv) with $\bar\delta_{\sup} \le \delta$ (cf. the proof of (i)), $\bar x^\infty \in S_\delta$. Hence $\lim_j d_{S_\delta}(\bar x^{k_j}) = 0$ by the continuity of d_{S_δ}, and thus $\overline{\lim}_k d_{S_\delta}(\bar x^k) = 0$.

(iv) Recalling (3.3) and (3.5), use $|x^{k+1}-x^k| \le \nu_k|g^k|$ and $\nu_k^2|g^k|^2 = 2\nu_k\gamma_k$. ∎

COROLLARY 4.2. *Suppose that the sequences $\{\nu_k\}$, $\{|g^k|\}$, and $\{\epsilon_k\}$ are bounded, and the extended objective f_S is coercive. Then the sequence $\{x^k\}$ is bounded.*

Proof. This follows from Theorem 4.1(ii), (iv) if $\sum_k \nu_k = \infty$. Otherwise, i.e., if $\sum_k \nu_k < \infty$, then by summing the inequality $|x^{k+1} - x^k| \le \nu_k|g^k|$ (cf. (3.3)) and using the assumption that $\sup_k |g^k| < \infty$ we get $\sum_k |x^{k+1} - x^k| < \infty$; hence $\{x^k\}$ converges. ∎

Remark 4.3.

(i) Theorem 4.1(ii) may be augmented as follows: (ii$_1$) *if $\delta = \sigma = 0$, then $S_*^\delta = S_\delta = S_*$ and $\lim_{k\to\infty} d_{S_*}(x^k) = 0$*; (ii$_2$) *if f is continuous on S, then $\overline{\lim}_{k\to\infty} f(x^k) \le \max_{S\cap S_*^\delta} f$ (so that $\lim_{k\to\infty} f(x^k) = f_*$ if $\delta = \sigma = 0$)*. Indeed, if $\delta = \sigma = 0$, then $S_\delta = S_*$ by (3.13), $\rho_\delta = 0$, and $S_*^\delta = S_*$ by (4.1), since S_* is closed, whereas if f is continuous on S, then by picking x^{k_j} such that $\lim_j f(x^{k_j}) = \overline{\lim}_k f(x^k)$ and $x^{k_j} \to x^\infty \in S_*^\delta$, from $x^{k_j} \in S$ we get $x^\infty \in S$ (since S is closed) and $\overline{\lim}_{k\to\infty} f(x^k) = f(x^\infty) \le \max_{S\cap S_*^\delta} f$.

(ii) Theorem 4.1(ii) subsumes [LPS00, Thm. 3] (where $\epsilon_k \to 0$, $\nu_k \to 0$, and $\nu_k|g^k|^2 \to 0$), a "stationary" version of [ShW96, Thm. 2.2] (where $\epsilon_k \downarrow 0$, $\nu_k|g^k|^2 \to 0$, $\sup_k \nu_k < \infty$ yield $\delta = \sigma = 0$), [Nur79, Thm. 2.8] (where S is bounded, $\delta = \sigma = 0$) and a convex version of [MGN87, Thm. 9.1] (where $S = \mathbb{R}^n$, $\epsilon_k \equiv 0$, $\nu_k \to 0$). Further, it subsumes [KiA91, Thm. 2] (where $\epsilon_k \to 0$, $\nu_k \to 0$, $\sup_k |g^k| < \infty$); the latter is a (mis)quotation of [NuZ77, Thm. 2], which, however, uses scaled stepsizes (cf. Remark 7.4(ii)).

5. Convergence when the iterates are bounded. We now show that the case where all the algorithmic quantities (i.e., x^k, g^k, ϵ_k, and ν_k) are bounded is analogous to the coercive case analyzed in Theorem 4.1. Only the statement of the following result is fairly complicated, since it does not presume that $S_* \ne \emptyset$.

THEOREM 5.1. *Suppose that $\sum_{k=1}^\infty \nu_k = \infty$, $\nu := \overline{\lim}_{k\to\infty} \nu_k < \infty$, $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, $\{x^k\}$ is bounded, and $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$. Then $\gamma := \overline{\lim}_{k\to\infty} \gamma_k \le \frac{1}{2}C^2\nu$, $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| \le C\nu$, and $\delta := \overline{\lim}_{k\to\infty} \delta_k \le \gamma + \epsilon$. For any $R \ge \underline{R} := \sup_k |x^k|$, consider the restricted problem*

$$f'_* := \inf f'_S \quad \textit{with} \quad f'_S := f_S + \mathrm{I}_{B_R}. \tag{5.1}$$

Let $S' := S \cap B_R$, $S'_* := \operatorname{Arg\,min} f'_S$, $S'_\delta := \{x : f'_S(x) \le f'_* + \delta\}$, and (cf. Lemma 2.4(i))

$$S_*^{\delta\prime} := S'_* + B_{\rho'_\delta+\sigma} \quad \textit{with} \quad \rho'_\delta := \max\left\{ d_{S'_*}(x) : x \in S'_\delta \right\}. \tag{5.2}$$

Then $f'_* \ge f_*$, $S'_* \supseteq S_*\cap B_R$, and $S'_\delta \supseteq S_\delta \cap B_R$, with equalities holding iff $S_*\cap B_R \ne \emptyset$. In fact, if S_* is nonempty and bounded, and R is large enough (e.g., $B_R \supset S_\delta$), then $f'_* = f_*$, $S'_* = S_*$, $S'_\delta = S_\delta$, $\rho'_\delta = \rho_\delta$ (cf. (4.1)), and $S_*^{\delta\prime} = S_*^\delta \cap B_R$. Moreover, we have the following statements:

(i) *$\underline{\lim}_{k\to\infty} d_{S_\delta}(x^k) = 0$ and $\{x^k\}$ has a cluster point in S_δ. Further, the assertions of Theorem* 3.2(ii)–(iii) *hold with $\bar\delta_{\sup} \le \delta$.*

(ii) *$\lim_{k\to\infty} d_{S_*^{\delta\prime}}(x^k) = 0$ and the cluster points of $\{x^k\}$ lie in $S_*^{\delta\prime}$.*

(iii) *$\{\bar x^k\}$ is bounded, its cluster points lie in S_δ, and $\lim_{k\to\infty} d_{S_\delta}(\bar x^k) = 0$.*

(iv) *If $\delta = 0$, then $S_* \ne \emptyset$, and $\lim_{k\to\infty} f(x^k) = f_*$ if f is continuous on S and $\sigma = 0$.*

Proof. By (5.1), f'_S is closed and convex (so are f_S and B_R), proper (since its domain $S' := S\cap B_R$ contains $\{x^k\}$ by the choice of R), and coercive (S' is bounded), so its optimal set $S'_* \subset B_R$ is nonempty and bounded. Of course, $f'_S \geq f_S$ and f'_S coincides with f_S on B_R. Hence $f'_* \geq f_*$, $S'_* \supseteq S_* \cap B_R$, and $S'_\delta \supseteq S_\delta \cap B_R$ (cf. (3.13)), with equalities holding throughout iff $S_* \cap B_R \neq \emptyset$. Indeed, if $f'_* = f_*$, then $\emptyset \neq S'_* \subset S_* \cap B_R$ and $S'_\delta \subset S_\delta \cap B_R$ from $f'_* + \delta < \infty$; conversely, if $f_S(x) = f_*$ for some $x \in B_R$, then $f_* = f'_S(x) \geq f'_* \geq f_*$ implies $f'_* = f_*$. Similarly, if S_* is nonempty and bounded, then, since S_δ is bounded, we may choose R such that $B_R \supset S_\delta \supset S_*$, in which case $S'_* = S_* \cap B_R = S_*$ and $S'_\delta = S_\delta \cap B_R = S_\delta$, so that $\rho'_\delta = \rho_\delta$ and $S^{\delta\prime}_* = S^\delta_* \cap B_R$ by (4.1) and (5.2).

Next, we may replace S and f_S in (1.2) by $S' := S \cap B_R$ and f'_S, since $\{x^k\} \subset S'$, whereas $g^k \in \partial_{\epsilon_k} f_S(x^k)$ implies $g^k \in \partial_{\epsilon_k} f'_S(x^k)$, using $f'_S(x^k) = f_S(x^k)$ and $f'_S \geq f_S$. Thus the algorithm works as if applied to problem (5.1), for which the assumptions of Theorem 4.1 hold with S_* replaced by S'_* (since $\nu < \infty$ and $C < \infty$). Therefore, the conclusions of Theorems 4.1 and 3.2(ii)–(iii) are valid with f_S replaced by f'_S, f_* by f'_*, etc. In particular, assertion (ii) follows from Theorem 4.1(ii), whereas Theorem 4.1(i), (iii) implies the first part of (i) as well as (iii) with S_δ replaced by S'_δ. For proving (i), (iii), and (iv), note that x^k and $f_S(x^k) = f'_S(x^k)$ are *independent* of R, for $R \geq \underline{R}$.

(i) Theorem 3.2(ii), (vi) with $\bar\delta_{\sup} \leq \delta$ gives $\underline{\lim}_k f_S(x^k) \leq f'_* + \delta$, using $f_S(x^k) = f'_S(x^k)$. Pick a subsequence $\{x^{k_j}\}$ such that $\lim_j f_S(x^{k_j}) = \underline{\lim}_k f_S(x^k)$. Since $\{x^{k_j}\}$ is bounded, we may assume that $x^{k_j} \to x^\infty$. Then by the closedness of f_S, $f_S(x^\infty) \leq f'_* + \delta$. Hence $f_S(x^\infty) \leq f_* + \delta$, since (cf. (5.1)) we can make f'_* arbitrarily close to f_* by increasing R. Thus $x^\infty \in S_\delta$ (cf. (3.13)), so $d_{S_\delta}(x^{k_j}) \to d_{S_\delta}(x^\infty) = 0$. By a similar argument, the assertions of Theorem 3.2(ii)–(iii) hold both with f_* replaced by f'_* and in their original form.

(iii) Since $\{x^k\} \subset B_R$, $\{\bar x^k\} \subset B_R$ by (3.9). Pick $\bar x^{k_j}$ such that $\lim_j d_{S_\delta}(\bar x^{k_j}) = \overline{\lim}_k d_{S_\delta}(\bar x^k)$. Extracting a subsequence, if necessary, suppose $\bar x^{k_j} \to \bar x^\infty$. As in the proof of (i), invoking Theorem 3.2(iv) with $\bar\delta_{\sup} \leq \delta$ we get $\bar x^\infty \in S'_\delta$ and then $\bar x^\infty \in S_\delta$. Hence $\lim_j d_{S_\delta}(\bar x^{k_j}) = 0$ by the continuity of d_{S_δ}, and thus $\overline{\lim}_k d_{S_\delta}(\bar x^k) = 0$.

(iv) If $\delta = 0$, then in the proof of (i) we have $x^\infty \in S_0 = S_*$ (cf. (3.13)), i.e., $S_* \neq \emptyset$. If additionally $\sigma = 0$ and f is continuous on S, then $\lim_k f(x^k) = f'_*$ by (ii) (cf. Remark 5.2(i) below), with $f'_* = f_*$ for R large enough so that $S_* \cap B_R \neq \emptyset$. □

Remark 5.2.

(i) Theorem 5.1(ii) may be augmented as follows: (ii$_1$) *if* $\delta = \sigma = 0$ (*e.g.*, $\nu = \epsilon = 0$), *then* $S^{\delta\prime}_* = S'_\delta = S'_*$ *and* $\lim_{k\to\infty} d_{S_*}(x^k) = 0$; (ii$_2$) *if* f *is continuous on* S, *then* $\overline{\lim}_{k\to\infty} f(x^k) \leq \max_{S\cap S^{\delta\prime}_*} f'_S$ (*so that* $\lim_{k\to\infty} f(x^k) = f'_*$ *if* $\delta = \sigma = 0$). Indeed, this follows as in Remark 4.3(i).

(ii) For $S = \mathbb{R}^n$, Theorem 5.1(i)–(ii) subsumes [Nur91, Thms. 2.3 and 2.4] and the results of [Nur82, sect. 6] (where $\nu_k \to 0$, either $\epsilon_k \to 0$ or $\epsilon_k \equiv \epsilon > 0$, $S_* \neq \emptyset$ is assumed *implicitly*, and the proofs are more complicated).

6. Bounding strategies. Our further results require the following definition.

DEFINITION 6.1. *We say that the algorithm employs a* locally bounded oracle *if* $g^k = g(x^k, \epsilon_k)$ *for all* k, *where the mapping* $S \times \mathbb{R}_+ \ni (x,\epsilon) \mapsto g(x,\epsilon) \in \partial_\epsilon f_S(x)$ *is locally bounded* (*bounded on bounded subsets of its domain*).

This concept is quite natural in view of the following comments.

Remark 6.2.

(i) In most applications, one has an oracle (*black box*) that, given $(x,\epsilon) \in S \times \mathbb{R}_+$, delivers an approximate subgradient $g_f(x,\epsilon) \in \partial_\epsilon f(x)$. Recall that for a fixed ϵ, $\partial_\epsilon f(\cdot)$ is locally bounded on S if f is finite on a neighborhood of S, in which case $\partial_\epsilon f(S)$ is bounded if S is bounded; also $\partial_\epsilon f(S)$ is bounded if f is finite-valued and polyhedral [HUL93, sect. XI.4.1]. In such cases one may use $g := g_f$, since $\partial_\epsilon f(\cdot) \subset \partial_\epsilon f_S(\cdot)$ on S. For some applications [KLL99a, sect. 9.4] one may choose a locally bounded g_f even when $\partial_\epsilon f(\cdot)$ is unbounded.

(ii) To handle the constraint $x \in S$ more efficiently, one may use the subgradient projection techniques of [KiU93], [Kiw96a, sect. 7], and [LPS96, sect. 3]. Thus, for $g_f(x,\epsilon) \in \partial_\epsilon f(x)$, we may let $g(x,\epsilon)$ be the projection of $g_f(x,\epsilon)$ onto the negative of the tangent cone of S at x so that $-g(x,\epsilon)$ is a feasible direction when S is polyhedral; e.g., for $S := \mathbb{R}^n_+$, $g(x,\epsilon)_j = \min\{g_f(x,\epsilon)_j, 0\}$ if $x_j = 0$, $g_f(x,\epsilon)_j$ otherwise. Then $g(x,\epsilon) \in \partial_\epsilon f_S(x)$, and the crucial property $|g(x,\epsilon)| \le |g_f(x,\epsilon)|$ ensures that g is locally bounded if g_f is bounded.

(iii) Note that if a locally bounded oracle is available, then f must be locally Lipschitz continuous on S [KLL99b, Rem. 3.9(ii)].

Of course, for a locally bounded oracle, $\{g^k\}$ is bounded if $\{x^k\}$ and $\{\epsilon_k\}$ are bounded. We now show that if the algorithm starts from any point in a fixed bounded trench of f_S and employs sufficiently small stepsizes and subgradient errors, then $\{x^k\}$ is bounded.

THEOREM 6.3. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Fix any point $\bar{x} \in S$ and a bounding tolerance $\bar{\delta} \in (0,\infty)$. Then there exist stepsize and error thresholds $\bar{\nu}_{\max} > 0$ and $\bar{\epsilon}_{\max} > 0$ with the following property: If the algorithm starts from a point $x^1 \in T_{f(\bar{x})}$ (e.g., $x^1 = \bar{x}$) and employs stepsizes $\nu_k \le \bar{\nu}_{\max}$ and errors $\epsilon_k \le \bar{\epsilon}_{\max}$ for all k, then $\{x^k\}$ stays in the bounded trench $T_{f(\bar{x})+\bar{\delta}}$ so that $\{g^k\}$ is bounded.*

Proof. Let $\beta := f(\bar{x})$, $\bar{\alpha} := \beta + \bar{\delta}$. Since the oracle is locally bounded, f_S is continuous on S (cf. Remark 6.2(iii)). By Lemma 2.4(ii), there exists $\bar{\rho} > 0$ such that $S \cap (T_\beta + B_{2\bar{\rho}}) \subset T_{\bar{\alpha}}$, whereas by Lemma 2.4(i) there is $\alpha > \beta$ such that $T^\alpha_\beta \subset T_\beta + B_{\bar{\rho}}$; thus

$$S \cap \left(T^\alpha_\beta + B_{\bar{\rho}} \right) \subset S \cap \left(T_\beta + B_{2\bar{\rho}} \right) \subset T_{\bar{\alpha}}. \tag{6.1}$$

Let

$$\bar{\epsilon}_{\max} := \tfrac{1}{2}(\alpha - \beta), \tag{6.2}$$

$$C := \sup\left\{ |g(x,\epsilon)| : x \in S \cap (T_\beta + B_{2\bar{\rho}}), \epsilon \le \bar{\epsilon}_{\max} \right\}, \tag{6.3}$$

$$\bar{\nu}_{\max} := \min\left\{ \bar{\rho}/C, (\alpha - \beta)/C^2 \right\}. \tag{6.4}$$

Note that $C < \infty$, since T_β is bounded and $\bar{\epsilon}_{\max} < \infty$.

Since $\{x^k\} \subset S$ and $f(x^1) \le f(\bar{x}) =: \beta$, we have $x^1 \in S \cap (T_\beta + B_{2\bar{\rho}})$.

Assuming $x^k \in S \cap (T_\beta + B_{2\bar{\rho}})$ for some $k \ge 1$, we now show that $x^{k+1} \in S \cap (T_\beta + B_{2\bar{\rho}})$. Using the bound $|x^{k+1} - x^k| \le \nu_k |g^k|$ (cf. (3.3)) with $|g^k| = |g(x^k, \epsilon_k)| \le C$ (cf. (6.3)) and $\nu_k \le \bar{\nu}_{\max} \le \bar{\rho}/C$ (cf. (6.4)) gives $|x^{k+1} - x^k| \le \bar{\rho}$. Hence if $x^k \in T_\alpha$, then from $T_\alpha \subset T^\alpha_\beta$ (cf. (2.2)), the first inclusion of (6.1), and the fact that $x^{k+1} \in S$ we get

$$x^{k+1} \in S \cap \left(x^k + B_{\bar{\rho}} \right) \subset S \cap \left(T_\alpha + B_{\bar{\rho}} \right) \subset S \cap \left(T^\alpha_\beta + B_{\bar{\rho}} \right) \subset S \cap \left(T_\beta + B_{2\bar{\rho}} \right).$$

Next, suppose $x^k \notin T_\alpha$, i.e.,

$$f(x^k) > \alpha. \tag{6.5}$$

Since $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, we have $|x^k - x| \le 2\bar\rho$ for $x = P_{T_\beta} x^k$. Next, by (6.2)–(6.4),

$$\epsilon_k \le \bar\epsilon_{\max} \le \tfrac{1}{2}(\alpha - \beta) \quad \text{and} \quad \tfrac{1}{2}|g^k|^2 \nu_k \le \tfrac{1}{2} C^2 \bar\nu_{\max} \le \tfrac{1}{2}(\alpha - \beta). \tag{6.6}$$

Using the estimate (3.1) with $f_S(x) \le \beta$ and the bounds (6.5) and (6.6), we obtain

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[\, f(x^k) - f(x) - \epsilon_k - \tfrac{1}{2}|g^k|^2 \nu_k \,\right] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le 2\bar\rho$ with $x \in T_\beta$, so $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$.

Therefore, by induction, for all k we have $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, and hence (cf. (6.3)) $|g^k| \le C$ and (cf. (6.1)) $x^k \in T_{\bar\alpha}$. ◻

In view of Theorem 6.3, we may employ the following *bounding strategy* that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$ and $\bar\delta > 0$, pick positive sequences $\{\nu^l_{\max}\}$ and $\{\epsilon^l_{\max}\}$ such that $\nu^l_{\max} \to 0$ and $\epsilon^l_{\max} \to 0$ if $l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu^l_{\max}$ and errors $\epsilon_k \le \epsilon^l_{\max}$ until for some k (if any) it is discovered that

$$f(x^k) > f(\bar x) + \bar\delta, \tag{6.7}$$

in which case increase l by 1, restart the algorithm, etc.

A *special case* of the above strategy consists of picking sequences $\nu_k \to 0$ and $\epsilon_k \to 0$, and resetting x^{k+1} to $\bar x$ (or the best point found so far) if (6.7) holds. Ensuring that $\sup_k |g^k| < \infty$, this version meets the assumptions of Theorem 4.1 if $\sum_k \nu_k = \infty$ and of Theorem 3.4 if additionally $\sum_k \nu_k^2 < \infty$ and $\sum_k \nu_k \epsilon_k < \infty$. However, the general version allows us to satisfy the assumptions of Theorem 4.1 with $\overline{\lim}_k \nu_k > 0$ and $\overline{\lim}_k \epsilon_k > 0$.

To avoid calculating $f(x^k)$, the test (6.7) may be replaced by $|x^k| > R$ for R such that $T_{f(\bar x)+\bar\delta} \subset B_R$; this ensures the boundedness of $\{x^k\}$ and $\{g^k\}$ as before. However, finding such R may be difficult, so the following result motivates an alternative bounding strategy.

THEOREM 6.4. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Then for each $\beta \in (f_*, \infty)$ and $\bar\epsilon_{\max} \in [0, \infty)$ there exists $\bar\nu_{\max} > 0$ such that if $f_S(x^1) \le \beta$, $\nu_k \le \bar\nu_{\max}$, and $\epsilon_k \le \bar\epsilon_{\max}$ for all k, then $\{x^k\}$ and $\{g^k\}$ are bounded.*

Proof. We show only how to modify the proof of Theorem 6.3. Let $\bar\alpha := \infty$, $\alpha > \beta + 2\bar\epsilon_{\max}$. Invoking Lemma 2.4(i), pick $\bar\rho > 0$ such that $T^\alpha_\beta \subset T_\beta + B_{\bar\rho}$. Then we have (6.1), whereas (6.2) is replaced by $\bar\epsilon_{\max} \le \frac{1}{2}(\alpha - \beta)$; the rest goes on as before. ◻

In view of Theorem 6.4, we may use the following bounding strategy that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$ and $\bar\epsilon_{\max} \ge 0$, pick positive sequences $\nu^l_{\max} \to 0$ and $R_l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu^l_{\max}$ and errors $\epsilon_k \le \bar\epsilon_{\max}$; if

$$|x^k| > R_l \tag{6.8}$$

for some k, then increase l by 1, restart the algorithm, etc.

The test (6.8) may be replaced by $\max\{|x^k - x^1|, \nu_k |g^k|, |g^k|\} > R_l$.

This strategy also meets the assumptions of Theorem 4.1, if $\sum_k \nu_k = \infty$, and of Theorem 3.4 if additionally $\sum_k \nu_k^2 < \infty$ and $\sum_k \nu_k \epsilon_k < \infty$. Note that, in contrast with (6.7), its resetting test (6.8) does not require calculating $f(x^k)$.

Yet another bounding strategy stems from the following extension of Corollary 4.2.

THEOREM 6.5. *Suppose that* $\hat{\nu} := \sup_k \nu_k$, $\hat{\gamma} := \sup_k \gamma_k$, *and* $\hat{\epsilon} := \sup_k \epsilon_k$ *are finite and* f_S *is coercive. Then* $\{x^k\}$ *is bounded.*

Proof. We show only how to modify the proof of Theorem 6.3. Let $\beta := f(x^1)$, $\bar{\alpha} := \infty$, $\alpha > \beta + 2\max\{\hat{\epsilon}, \hat{\gamma}\}$. Invoking Lemma 2.4(i), pick $\bar{\rho} \geq (2\hat{\gamma}\hat{\nu})^{1/2}$ such that $T_\beta^\alpha \subset T_\beta + B_{\bar{\rho}}$. Then, by (3.3) and (3.5), we have $|x^{k+1} - x^k|^2 \leq \nu_k^2 |g^k|^2 = 2\nu_k\gamma_k$ and hence $|x^{k+1} - x^k| \leq \bar{\rho}$, $\epsilon_k \leq \frac{1}{2}(\alpha - \beta)$ and $\frac{1}{2}|g^k|^2\nu_k \leq \frac{1}{2}(\alpha - \beta)$ as in (6.6); the rest goes on as before. ∎

Theorem 6.5 suggests the following bounding strategy with resets indexed by $l = 1, 2, \ldots$. Fixing $\bar{x} \in S$, $\bar{\epsilon}_{\max} \in [0, \infty)$, and $\gamma_{\max} \in (0, \infty)$, pick a positive sequence $\nu_{\max}^l \to 0$. For the current $l \geq 1$, start the algorithm from $\bar{x}$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \leq \nu_{\max}^l$ and errors $\epsilon_k \leq \bar{\epsilon}_{\max}$; if $\gamma_k > \gamma_{\max}$ for some k, then increase l by 1, restart the algorithm, etc. Under the assumptions of Theorem 6.4, only finitely many resets occur (otherwise we would have $\hat{G} := \sup_k |g^k| < \infty$ and $\frac{1}{2}\hat{G}^2\nu_{\max}^l > \gamma_{\max}$ at each reset, contradicting $\nu_{\max}^l \to 0$), so Theorem 6.5 implies the boundedness of $\{x^k\}$. (A special case of this strategy consists of using sequences $\nu_k \to 0$ and $\epsilon_k \leq \bar{\epsilon}_{\max}$, and resetting x^{k+1} to x^1 whenever $\gamma_k > \gamma_{\max}$.) Alternatively, the test $\gamma_k > \gamma_{\max}$ may be replaced by $|g^k| > G_l$, where $G_l \to \infty$ as $l \to \infty$ (e.g., $G_{l+1} := \max\{|g^k|, 10G_l\}$).

Remark 6.6. For $S = \mathbb{R}^n$ and $\epsilon_k \equiv 0$, Theorem 6.3 subsumes in the convex case [MGN87, Lem. 9.1] (which employs (6.7) with $\bar{x} = x^1$), whereas Theorem 6.4 subsumes a result of [Sho79, p. 39]. We note that the proof of [MGN87, Lem. 9.1] is quite complicated, whereas that of [Sho79, p. 39] does not extend to the constrained case.

7. Using scaled stepsizes.

7.1. Extension of Ermoliev's framework. We now highlight an idea that is implicit in the pioneering paper of Ermoliev [Erm66, sect. 9]: to ensure convergence, the stepsize ν_k may be chosen as $\nu_k := \lambda_k\mu_k$, where λ_k is fairly arbitrary (e.g., $\lambda_k := k^{-1}$), but μ_k should damp the possible growth of $|g^k|$. We first discuss general conditions on the choice of μ_k and then provide several examples.

THEOREM 7.1. *Suppose that* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$ *and the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, *and* $\mu_k > 0$ *such that*

$$\bar{\gamma} := \overline{\lim_{k\to\infty}}\, \tfrac{1}{2}\mu_k|g^k|^2 < \infty, \tag{7.1}$$

$$\underline{\lim_{k\to\infty}}\, \mu_k > 0 \quad \textit{whenever} \quad \{x^k\} \quad \textit{is bounded.} \tag{7.2}$$

Then $\sum_{k=1}^\infty \nu_k = \infty$ *whenever* $\{x^k\}$ *is bounded. Further, we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) \leq f_* + \delta$, *where* $\delta := \overline{\lim}_{k\to\infty} \delta_k \leq \gamma + \epsilon$ *with* $\gamma := \overline{\lim}_{k\to\infty} \gamma_k \leq \bar{\gamma}\lambda$.

(ii) *If* f_S *is coercive and* $\bar{\sigma} := \overline{\lim}_{k\to\infty} \nu_k|g^k|$ *is finite, which holds if*

$$\overline{\lim_{k\to\infty}}\, \mu_k|g^k| < \infty \quad \textit{or} \quad \mu := \overline{\lim_{k\to\infty}}\, \mu_k < \infty, \tag{7.3}$$

then the conclusions of Theorem 4.1 *hold with* $\nu := \overline{\lim}_{k\to\infty} \nu_k \le \lambda\mu$ *and*

$$\sigma := \overline{\lim_{k\to\infty}} |x^{k+1} - x^k| \le \bar\sigma \le \lambda \min\left\{ \overline{\lim_{k\to\infty}} \mu_k |g^k|, (2\mu\bar\gamma)^{1/2} \right\}. \tag{7.4}$$

(iii) *If additionally* $\sum_{k=1}^\infty \lambda_k^2 < \infty$ *and the assumptions* $\epsilon < \infty$ *and* $\bar\gamma < \infty$ *are replaced by* $\sum_{k=1}^\infty \nu_k \epsilon_k < \infty$ *and* $\sup_k \mu_k |g^k| < \infty$ (*retaining* $\sum_{k=1}^\infty \lambda_k = \infty$ *and* (7.2)) *then we have the following statements:*

(iii$_1$) $\underline{\lim}_{k\to\infty} f(x^k) = f_*$.

(iii$_2$) $S_* \ne \emptyset$ *iff* $\{x^k\}$ *is bounded.*

(iii$_3$) *If* $S_* \ne \emptyset$, *then the assumptions of Theorem* 3.4 *hold*; *in particular,* $\{x^k\}$ *and* $\{\bar x^k\}$ *converge to some* $x^\infty \in S_*$.

Proof. Note that $\sum_k \lambda_k = \infty$ and (7.2) imply $\sum_k \nu_k = \infty$ whenever $\{x^k\}$ is bounded.

(i) For contradiction, suppose there exist $x \in S$, $v > 0$, and k_v such that $f(x^k) \ge f(x) + \delta + v$ for all $k \ge k_v$. Pick $k_v' \ge k_v$ such that $\delta_k \le \delta + v$ for all $k \ge k_v'$. Then (3.6) yields $|x^{k+1} - x| \le |x^k - x|$ for all $k \ge k_v'$. Thus $\{x^k\}$ is bounded, so $\sum_k \nu_k = \infty$. Hence Theorem 3.2(ii), (vi) gives $\bar\delta_{\sup} \le \delta$ and $\underline{\lim}_k f(x^k) \le f_* + \delta$, a contradiction.

(ii) We have $\sigma \le \bar\sigma < \infty$ from $|x^{k+1} - x^k| \le \nu_k |g^k|$ (cf. (3.3)), $\bar\sigma \le \lambda\, \overline{\lim}_k \mu_k |g^k|$, and $\bar\sigma^2 \le \lambda^2 \mu 2\bar\gamma$ by the definitions of $\bar\sigma$, ν_k, λ, $\bar\gamma$, and μ. Using (i) in the proof of Theorem 4.1(i) gives $\underline{\lim}_k d_{S_\delta}(x^k) = 0$. Then the proof of Theorem 4.1(ii) yields the boundedness of $\{x^k\}$, so $\sum_k \nu_k = \infty$. Hence we may invoke Theorem 3.2(ii), (vi) in the proof of Theorem 4.1(i), and Theorem 3.2(iv) in the proof of Theorem 4.1(iii).

(iii) Since $\tilde C := \sup_k \mu_k |g^k| < \infty$, we have $\sum_k \nu_k^2 |g^k|^2 \le \tilde C^2 \sum_k \lambda_k^2 < \infty$. (iii$_1$) Suppose $\underline{\lim}_k f(x^k) > f_*$. Thus there are $x \in S$ and $\bar k$ such that $f(x^k) \ge f(x)$ for all $k \ge \bar k$. Then by the proof of "(i) $\Rightarrow$ (ii)" in Theorem 3.4, $\{x^k\}$ is bounded, so $\sum_k \nu_k = \infty$ and Theorems 3.4 and 3.2(iii) yield $\underline{\lim}_k f(x^k) = f_*$, a contradiction. (iii$_2$–iii$_3$) If $S_* \ne \emptyset$, then $\{x^k\}$ is bounded by Theorem 3.4. On the other hand, if $\{x^k\}$ is bounded, then $\sum_k \nu_k = \infty$, so the conclusion follows from Theorem 3.4. ☐

Remark 7.2. When $\sup_k \epsilon_k < \infty$, (7.2) holds if the oracle is locally bounded and

$$\underline{\lim_{k\to\infty}}\, \mu_k > 0 \quad \text{whenever} \quad \{g^k\} \quad \text{is bounded.} \tag{7.5}$$

Next, we exhibit several choices of the scaling coefficients μ_k for Theorem 7.1 that ensure convergence without *any* indirect assumptions on the boundedness of $\{g^k\}$ which are implicit in the results of sections 3 and 4, and hence do not need the bounding techniques of section 6.

Example 7.3. For a locally bounded oracle (with $\sup_k \epsilon_k < \infty$) and a constant $G > 0$, the requirements (7.1) and (7.3) of Theorem 7.1 and (7.5) are met by the scaling coefficients

$$\mu_k := \max\left\{ |g^k|, |g^k|^2/G \right\}^{-1} = \min\left\{ 1, G/|g^k| \right\} |g^k|^{-1}, \tag{7.6}$$

where G replaces $|g^k|$ if $|g^k| = 0$ (with $\mu_k |g^k|^2 \le G$, $\mu_k |g^k| \le 1$),

$$\mu_k := \max\left\{ 1, |g^k|^2/G^2 \right\}^{-1} = \min\left\{ 1, G^2/|g^k|^2 \right\} \tag{7.7}$$

(with $\mu_k |g^k|^2 \le G^2$, $\mu_k |g^k| \le G$), and

$$\mu_k := \max\left\{ G^2, |g^k|^2 \right\}^{-1} = \min\left\{ 1, G^2/|g^k|^2 \right\} G^{-2} \tag{7.8}$$

(with $\mu_k|g^k|^2 \le 1$, $\mu_k|g^k| \le G^{-1}$); yet another choice of [NuZ77, Thm. 2] with $G \ge 1$ is

$$\mu_k := \begin{cases} 1 & \text{if } |g^k| \le G, \\ |g^k|^{-2} & \text{otherwise.} \end{cases} \tag{7.9}$$

The requirements (7.5), (7.3), and $\sup_k \mu_k|g^k| < \infty$ of Theorem 7.1(iii) are met by

$$\mu_k := |g^k|^{-1}, \tag{7.10}$$

the classical scaling of Shor [Sho62], and its popular variants

$$\mu_k := \big(G + |g^k| \big)^{-1},\ \mu_k := \max\big\{ G, |g^k| \big\}^{-1},\ \text{or}\ \mu_k := \big(G^2 + |g^k|^2 \big)^{-1/2} \tag{7.11}$$

(with $\mu_k|g^k| \le 1$), as well as by the choice of [Lis86]

$$\mu_k := \max\big\{ \lambda_k, |g^k| \big\}^{-1} = \min\big\{ \lambda_k^{-1}, |g^k|^{-1} \big\} \tag{7.12}$$

(using $\sup_k \lambda_k < \infty$ for (7.5)); note that if $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$ (e.g., $\{x^k\}$ is bounded), then also (7.1) holds with $\bar\gamma \le \frac{1}{2}C$, as required in Theorem 7.1(i)–(ii) (and $\bar\sigma \le \lambda$ in (7.4)). Next,

$$\mu_k := |g^k|^{-2} \tag{7.13}$$

satisfies (7.1) (with $\bar\gamma \le 1/2$) and (7.5) as required in Theorem 7.1(i), as well as (7.3) if $\underline{\lim}_k |g^k| > 0$ (which typically holds in the nondifferentiable case). Thus (7.8) with a "small" G may be regarded as a regularized version of (7.13) that ensures (7.3), but

$$\mu_k := \max\big\{ \lambda_k^2, |g^k|^2 \big\}^{-1} \tag{7.14}$$

also meets the requirements of Theorem 7.1(i)–(ii) (with $\bar\gamma \le 1/2$, $\nu_k|g^k| \le 1$, $\bar\sigma \le 1$). Note that (7.6)–(7.11) may use a variable $G = G_k \in [G_{\min}, G_{\max}] \subset (0, \infty)$.

Remark 7.4.

(i) Theorem 7.1(i) and its proof correct the proof of [Erm66, sect. 9], where the assumption (7.2) was *implicit* (and the claim that $f(x^k) \to f_*$ was *not* proved). Equation (7.2) is also implicit in [Erm76, Thm. I.3.5] (where $\sup_k \mu_k|g^k| < \infty$ should be replaced by (7.1)) and in [Erm76, Thm. I.3.6] (where $\sup_k \mu_k < \infty$ is implicit); the latter is subsumed by Theorem 7.1(ii). Theorem 7.1(iii$_3$) subsumes [Erm76, Thm. III.1.4] (in the deterministic case).

(ii) Theorem 7.1(ii) subsumes [NuZ77, Thm. 2], which uses (7.9) and $\epsilon = \lambda = 0$. Theorem 7.1(iii) subsumes [Sch83, Lem. on p. 539] with $\mu_k := (G^2 + |g^k|^2)^{-1/2}$ and $\epsilon_k \equiv 0$, and [AIS98, Thm. 1], in which $\mu_k := \max\{1, |g^k|\}^{-1}$ and $\epsilon_k \le C_\epsilon \lambda_k$ with $C_\epsilon < \infty$. Theorem 7.1(iii$_1$) subsumes [Lis86, Thm. on p. 70], which uses (7.12) and $\epsilon_k \equiv 0$, whereas Theorem 7.1(iii$_3$) subsumes [LPS00, Thm. 10] (with $\mu_k := \max\{1, |g^k|\}^{-1}$, $\sum_k \lambda_k \epsilon_k < \infty$, $\epsilon_k \to 0$) and [DeV81, Thm. III.4.5], which uses (7.10) and $\epsilon_k \equiv 0$.

We also have an analogue of Theorem 5.1 for scaled stepsizes.

THEOREM 7.5. *Assume that $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, $\{x^k\}$ is bounded, and $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$ (e.g., the oracle is locally bounded). Suppose that the algorithm employs stepsizes $\nu_k := \lambda_k \mu_k$ with $\lambda_k, \mu_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, $\underline{\lim}_{k\to\infty} \mu_k > 0$, such that $\bar\gamma := \overline{\lim}_{k\to\infty} \frac{1}{2}\mu_k|g^k|^2 < \infty$ and $\bar\sigma := \overline{\lim}_{k\to\infty} \nu_k|g^k| < \infty$.*

Let $\mu := \overline{\lim}_{k\to\infty} \mu_k$ *and* $\nu := \overline{\lim}_{k\to\infty} \nu_k$. *Then the conclusions of Theorem* 5.1 *hold with* $\gamma \le \bar\gamma\lambda$, $\bar\gamma \le \frac{1}{2}C^2\mu$, $\sigma \le \bar\sigma \le \lambda \min\{\overline{\lim}_{k\to\infty} \mu_k|g^k|, (2\mu\bar\gamma)^{1/2}\}$, *and* $\nu \le \lambda\mu$.

Proof. Invoke Theorem 7.1(ii) in the proof of Theorem 5.1. □

Remark 7.6.

(i) For a locally bounded oracle, the requirements of Theorem 7.5 are met by the scaling coefficients given by (7.6)–(7.12).

(ii) Theorem 7.5 subsumes [MGN87, Thm. 9.2] in the convex case with $\lambda = \epsilon = 0$.

7.2. Analysis of Shor-type scalings. Additional results for the Shor-type scalings (7.10)–(7.12) require the following assumption.

Assumption 7.7. The objective f is finite-valued and $g^k \in \partial_{\epsilon_k} f(x^k)$ for all k.

Under Assumption 7.7, the objective f is continuous, as required for the following basic estimates inspired by [Nes84, Lem. 1].

LEMMA 7.8. *Suppose Assumption* 7.7 *holds. Fixing a point* $x \in S$, *define the function*

$$\omega_x(\rho) := \max_{x+B_\rho} f \quad \text{for} \quad \rho \ge 0, \tag{7.15}$$

and let ρ_k^+ *be the distance from the point* x *to the halfspace* $\{y : \langle g^k, x^k - y\rangle \le 0\}$:

$$\rho_k^+ := \max\{\rho_k, 0\} \quad \text{with} \quad \rho_k := \begin{cases} \langle g^k/|g^k|, x^k - x\rangle & \text{if } g^k \ne 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7.16}$$

The function ω_x *is continuous and nondecreasing, and we have the estimate*

$$f(x^k) \le \omega_x(\rho_k^+) + \epsilon_k. \tag{7.17}$$

The stepsize $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$ *and* $\mu_k \le |g^k|^{-1}$ (*as in* (7.10)–(7.12)) *produces*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k|g^k|\left(\rho_k - \tfrac{1}{2}\nu_k|g^k|\right) \le -2\lambda_k\mu_k|g^k|\left(\rho_k - \tfrac{1}{2}\lambda_k\right). \tag{7.18}$$

Proof. Suppose $f(x) < f(x^k) - \epsilon_k$. (Otherwise (7.17) holds with $\omega_x(\rho_k^+) \ge f(x)$.) Then $\rho_k > 0$ (since $g^k \in \partial_{\epsilon_k} f(x^k)$). The point $\hat x := x + \frac{\rho_k}{|g^k|}g^k$ satisfies $|\hat x - x| = \rho_k$ and $\langle g^k, x^k - \hat x\rangle = 0$, so $f(\hat x) \le \omega_x(\rho_k)$ and $f(\hat x) \ge f(x^k) - \epsilon_k$ (from $g^k \in \partial_{\epsilon_k} f(x^k)$); thus (7.17) holds. For (7.18), rewrite (3.4) with $\nu_k := \lambda_k\mu_k$ and use $\mu_k|g^k| \le 1$. □

We have the following analogue of Theorem 7.1(i) for the scalings (7.10)–(7.12).

THEOREM 7.9. *Suppose Assumption* 7.7 *holds,* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, *and the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, *and* μ_k *chosen as in* (7.10)–(7.12). *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) \le \inf_{x\in S} \max_{x+B_{\lambda/2}} f + \epsilon$.

(ii) *If* $\lambda = 0$ (*i.e.*, $\lim_{k\to\infty} \lambda_k = 0$), *then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \epsilon$.

(iii) *If* $S_* \ne \emptyset$, *then* $\underline{\lim}_{k\to\infty} f(x^k) \le \inf_{x\in S_*} \max_{x+B_{\lambda/2}} f + \epsilon \le \sup_{S_*+B_{\lambda/2}} f + \epsilon$.

Proof. We need only to prove item (i), since (ii) and (iii) follow immediately from (i).

First, suppose μ_k is chosen via (7.10). Then for $x \in S$ and ρ_k defined by (7.16) we have

$$\underline{\lim}_{k\to\infty} \rho_k \le \tfrac{1}{2}\lambda. \tag{7.19}$$

Indeed, summing up (7.18) with $\mu_k|g^k|$ replaced by 1 produces the Cesáro estimate

$$\bar{\rho}_k := \frac{\sum_{j=1}^k \lambda_j \rho_j}{\sum_{j=1}^k \lambda_j} \le \frac{|x^1 - x|^2 + \sum_{j=1}^k \lambda_j^2}{2\sum_{j=1}^k \lambda_j}, \tag{7.20}$$

which combined with $\sum_k \lambda_k = \infty$ yields $\underline{\lim}_k \rho_k \le \overline{\lim}_k \bar{\rho}_k \le \frac{1}{2}\lambda$ (cf. Lemma 2.1). By (7.17) and (7.19), we have $\underline{\lim}_{k\to\infty} f(x^k) \le \max_{x+B_{\lambda/2}} f + \epsilon$ for each $x \in S$, as required.

Similarly, for the remaining choices (7.11)–(7.12), assertion (i) is established if (7.19) holds, so suppose $\underline{\lim}_k \rho_k > \frac{1}{2}\lambda$ for some $x \in S$. Thus, since $\lambda := \overline{\lim}_k \lambda_k$, we have $\rho_k > \frac{1}{2}\lambda_k$ for large k and (7.18) shows that $\{x^k\}$ is bounded. We consider two cases.

First, suppose $\underline{\lim}_k |g^k| = 0$. Then a subsequence $g^{k_j} \to 0$, and taking limits in the subgradient inequality $f(y) \ge f(x^{k_j}) - \epsilon_{k_j} + \langle g^{k_j}, y - x^{k_j}\rangle$ gives $\underline{\lim}_k f(x^k) \le f(y) + \epsilon$ for each y; thus assertion (i) holds.

Second, suppose $\underline{\lim}_k |g^k| > 0$. Write $\nu_k := \lambda_k\mu_k$ as $\nu_k = \hat{\lambda}_k\hat{\mu}_k$ with $\hat{\lambda}_k := \lambda_k\mu_k|g^k|$ and $\hat{\mu}_k := |g^k|^{-1}$. Note that $\hat{\lambda}_k \le \lambda_k$ (since $\mu_k \le |g^k|^{-1}$) and $\underline{\lim}_k \mu_k|g^k| > 0$ for the choices (7.11)–(7.12) (using $\underline{\lim}_k |g^k| > 0$ and $\overline{\lim}_k \lambda_k < \infty$ for (7.12)). The first property gives $\hat{\lambda} := \overline{\lim}_k \hat{\lambda} \le \lambda$, whereas the second one combined with $\sum_k \lambda_k = \infty$ implies $\sum_k \hat{\lambda}_k = \infty$. Hence by replacing λ_k, μ_k by $\hat{\lambda}_k$, $\hat{\mu}_k$ in the argument of the first paragraph we obtain assertion (i) with λ replaced by $\hat{\lambda}$; since $\hat{\lambda} \le \lambda$, (i) must hold for λ as well. □

A result on finite convergence is given in part (ii) of the following corollary.

Corollary 7.10. *Under the assumptions of Theorem* 7.9*, suppose that the optimal set* S_* *is nonempty and* $\epsilon_k \equiv 0$ *so that* $\lambda := \overline{\lim}_{k\to\infty} \lambda_k$ *determines the asymptotic accuracy. Then we have the following statements:*

(i) *For every* $\hat{\delta} > 0$, *if* λ *is small enough so that* $\omega_x(\frac{1}{2}\lambda) < f_* + \hat{\delta}$ *for some* $x \in S_*$ *(cf.* (7.15)*), then* $\underline{\lim}_{k\to\infty} f(x^k) < f_* + \hat{\delta}$.

(ii) *For every* $\rho > \frac{1}{2}\lambda$ *and* $x \in S_*$, *if* $\omega_x(\rho) > f_*$ *or the Shor scaling* (7.10) *is used, then there is an iteration* $\hat{k}$ *such that* $f(x^{\hat{k}}) = f(\hat{x})$ *for a point* $\hat{x}$ *satisfying* $|\hat{x} - x| < \rho$; *in particular, if* $x + B_\rho \subset S_*$ *and the Shor scaling* (7.10) *is employed, then* $x^{\hat{k}} \in S_*$.

Proof. (i) By (7.15) and Theorem 7.9(iii), $\underline{\lim}_k f(x^k) \le \omega_x(\frac{1}{2}\lambda)$.

(ii) The function ω_x is increasing for ρ such that $\omega_x(\rho) > f(x) = f_*$ (since any maximizer y of (7.15) satisfies $|y - x| = \rho$ by convexity), so $\underline{\lim}_k f(x^k) \le \omega_x(\frac{1}{2}\lambda) < \omega_x(\rho)$ yields the existence of $\hat{k}$ such that $f(x^{\hat{k}}) < \omega_x(\rho)$. For the scaling (7.10), since $\underline{\lim}_k \rho_k \le \frac{1}{2}\lambda < \rho$ by (7.19), for $\hat{k}$ such that $\rho_{\hat{k}}^+ < \rho$ we have $f(x^{\hat{k}}) \le \omega_x(\rho_{\hat{k}}^+)$ by (7.17). The existence of $\hat{x}$ follows from the continuity of f in (7.15), with $f(\hat{x}) = f_*$ if $x + B_\rho \subset S_*$. □

The Shor-type scalings (7.10)–(7.12) have the following analogue of Theorem 7.1(ii).

Theorem 7.11. *Suppose Assumption* 7.7 *holds,* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, *the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, μ_k *chosen as in* (7.10)–(7.12)*, and* f_S *is coercive. Then* $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| \le \lambda$. *Let*

$$\hat{\delta} := \hat{\gamma} + \epsilon \quad \textit{with} \quad \hat{\gamma} := \max_{S_* + B_{\lambda/2}} f - f_*. \tag{7.21}$$

Then we have the following statements:

(i) $\underline{\lim}_{k\to\infty} d_{S_{\hat\delta}}(x^k) = 0$ *and* $\{x^k\}$ *has a cluster point in* $S_{\hat\delta}$.

(ii) $\lim_{k\to\infty} d_{S_*^{\hat\delta}}(x^k) = 0$, *where* $S_*^{\hat\delta}$ *is the neighborhood of* S_* *defined by* (*cf. Lemma* 2.4(i))

$$S_*^{\hat\delta} := S_* + B_{\rho_{\hat\delta}+\sigma} \quad \textit{with} \quad \rho_{\hat\delta} := \max\left\{ d_{S_*}(x) : x \in S_{\hat\delta} \right\}. \tag{7.22}$$

Thus $\{x^k\}$ *is bounded and its cluster points belong to* $S_*^{\hat\delta}$.

(iii) $C := \overline{\lim}_{k\to\infty} |g^k|$ *is finite and the conclusions of Theorem* 7.1(i)–(ii) *hold with* $\bar\gamma \le \frac{1}{2}C$; *in particular, the conclusions of Theorem* 4.1 *hold with* $\gamma \le \frac{1}{2}C\lambda$ *and* $\sigma \le \lambda$ *so that assertions* (i) *and* (ii) *hold with* $\hat\delta$ *replaced by* $\min\{\delta, \hat\delta\}$, *where* $\delta := \overline{\lim}_k \delta_k \le \frac{1}{2}C\lambda + \epsilon$.

Proof. As in the proof of Theorem 4.1, the closedness and coercivity of f_S imply that the sets $S_* \subset S_{\hat\delta} \subset S_* + B_{\rho_{\hat\delta}} \subset S_*^{\hat\delta}$ are nonempty and compact (with $\hat\gamma < \infty$ because f is continuous). Further, (3.3) implies $|x^{k+1} - x^k| \le \lambda_k \mu_k |g^k| \le \lambda_k$, and hence $\sigma \le \lambda$.

(i) By Theorem 7.9(iii) and (7.21), we have $\underline{\lim}_k f(x^k) \le f_* + \hat\gamma + \epsilon = f_* + \hat\delta$, so the conclusion follows upon replacing δ by $\hat\delta$ in the proof of Theorem 4.1(i).

(ii) Fixing $v > 0$, let $\lambda_v := \lambda + v$, $\gamma_v := \max_{S_*+B_{\lambda_v/2}} f - f_*$, $\delta_v := \gamma_v + \epsilon + v$, $\alpha := \alpha_v := f_* + \delta_v$, $\rho_\alpha := \max_{T_\alpha} d_{S_{\hat\delta}}$ (so that $T_\alpha \subset S_{\hat\delta} + B_{\rho_\alpha}$; cf. (2.1), (2.2)), and (cf. (7.22))

$$V_v := S_*^{\hat\delta} + B_{\rho_\alpha+v} = S_* + B_{\rho_{\hat\delta}+\sigma+\rho_\alpha+v}. \tag{7.23}$$

By (7.21), $\gamma_v \ge \hat\gamma$, $\delta_v > \hat\delta$, and $\alpha_v > f_* + \hat\delta$. Since S_* is compact and f is continuous, for $v \downarrow 0$ we have $\gamma_v \downarrow \hat\gamma$, $\delta_v \downarrow \hat\delta$, $\alpha_v \downarrow f_* + \hat\delta$, and $\rho_\alpha \downarrow 0$ (cf. Lemma 2.4(i) with $\beta := f_* + \hat\delta$).

Since $\lambda := \overline{\lim}_k \lambda_k$, $\epsilon := \overline{\lim}_k \epsilon_k$ and $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$, there is $k_v < \infty$ such that

$$\lambda_k \le \lambda_v, \quad \epsilon_k \le \epsilon + v, \quad \text{and} \quad |x^{k+1} - x^k| \le \sigma + v \quad \forall k \ge k_v. \tag{7.24}$$

Since $\underline{\lim}_k d_{S_{\hat\delta}}(x^k) = 0$ by (i), there exists $k = k_v' \ge k_v$ such that $x^k \in S_{\hat\delta} + B_v$; then $S_{\hat\delta} \subset S_*^{\hat\delta}$ implies $x^k \in V_v$ (cf. (7.23)).

Assuming $x^k \in V_v$ for some $k \ge k_v'$, we now show that $x^{k+1} \in V_v$. If $x^k \in T_\alpha$, then from the third inequality of (7.24), $T_\alpha \subset S_{\hat\delta} + B_{\rho_\alpha}$, and $S_{\hat\delta} \subset S_* + B_{\rho_{\hat\delta}}$ (cf. (7.22)) we get

$$x^{k+1} \in T_\alpha + B_{\sigma+v} \subset S_{\hat\delta} + B_{\rho_\alpha+\sigma+v} \subset S_* + B_{\rho_{\hat\delta}} + B_{\rho_\alpha+\sigma+v} = S_* + B_{\rho_{\hat\delta}+\sigma+\rho_\alpha+v},$$

so $x^{k+1} \in V_v$ (cf. (7.23)). Thus suppose $x^k \notin T_\alpha$. Then, by the second inequality of (7.24),

$$f(x^k) - \epsilon_k > \alpha - \epsilon_k = f_* + \gamma_v + \epsilon + v - \epsilon_k \ge f_* + \gamma_v = \max_{S_*+B_{\lambda_v/2}} f,$$

so for $x = P_{S_*} x^k$, by Lemma 7.8, we have $\omega_x(\frac{1}{2}\lambda_v) < f(x^k) - \epsilon_k \le \omega_x(\rho_k^+)$, $\rho_k > \frac{1}{2}\lambda_v$, and $|x^{k+1} - x| \le |x^k - x|$ because $\lambda_k \le \lambda_v$ in (7.18) due to the first inequality of (7.24). Since $x \in S_*$ and $x^k \in V_v$, the inequality $|x^{k+1} - x| \le |x^k - x|$ and (7.23) yield $x^{k+1} \in V_v$.

Therefore, by induction for each $k \ge k'_v$, $x^k \in V_v$ and hence (cf. (7.23)) $d_{S_*^{\hat\delta}}(x^k) \le \rho_\alpha + v$. Since $\rho_\alpha \downarrow 0$ as $v \downarrow 0$, $d_{S_*^{\hat\delta}}(x^k) \to 0$. The rest follows as in the proof of Theorem 4.1(ii).

(iii) We have $\sup_k |g^k| < \infty$, since $\{x^k\}$ is bounded, $\sup_k \epsilon_k < \infty$, and the oracle is locally bounded under Assumption 7.7 (cf. Remark 6.2(i)). The conclusion follows from Theorem 7.1 and the discussion of (7.10)–(7.12) in Example 7.3. ∎

Remark 7.12.

(i) Theorem 7.11(ii) may be augmented as follows: (ii$_1$) *if* $\lambda = \epsilon = 0$, *then* $S_*^{\hat\delta} = S_{\hat\delta} = S_*$ *and* $\lim_{k\to\infty} d_{S_*}(x^k) = 0$; (ii$_2$) $\overline{\lim}_{k\to\infty} f(x^k) \le \max_{S\cap S_*^{\hat\delta}} f$ (*so that* $\lim_{k\to\infty} f(x^k) = f_*$ *if* $\lambda = \epsilon = 0$). Indeed, this follows as in Remark 4.3(i).

(ii) For $\lambda > 0$ (i.e., nonvanishing stepsizes), the asymptotic accuracy determined by $\hat\gamma$ in (7.21) may depend on the behavior of f outside the feasible set S, whereas the corresponding bound of Theorem 4.1 expressed by $\gamma \le \frac{1}{2}\lambda\,\overline{\lim}_k |g^k|$ depends on the properties of f seen by the algorithm inside S; the bound of Theorem 7.11(iii) using $\min\{\delta, \hat\delta\}$ combines the best of both worlds.

(iii) The estimate (7.17) extends [Nes84, Lem. 1] (to $\epsilon_k > 0$). Theorem 7.9 subsumes [Pol67, Thm. 1] (which uses (7.10) and $\epsilon_k \equiv 0$). For the Shor scaling (7.10), Corollary 7.10 subsumes [Sho79, Thm. 2.1 and Cors. 1–2] (where $\lambda_k \equiv \lambda > 0$) and [DeV81, Cor. III.4.1] (where $\lambda = 0$), whereas Theorem 7.11(i)–(ii) subsumes [DeV81, Thms. III.4.1–4] and some results of [DeV81, sect. IV.5]; the proof of a related result [LPS00, Thm. 6] is wrong.

7.3. Shor's bounding strategy. The following result helps in analyzing the bounding strategy of Shor [Sho79, Thm. 2.4].

PROPOSITION 7.13. *Suppose that Assumption* 7.7 *holds and* f_S *is coercive. Fix any point* $\bar x \in S$, *a step bound* $\bar\rho \in (0,\infty)$, *and an error threshold* $\bar\epsilon_{\max} \in [0,\infty)$. *If* $f_S(x^1) \le f(\bar x)$, $\nu_k|g^k| \le \bar\rho$, *and* $\epsilon_k \le \bar\epsilon_{\max}$ *for all* k, *then* $\{x^k\}$ *and* $\{g^k\}$ *are bounded.*

Proof. Let $\alpha := \max_{\bar x + B_{\bar\rho}} f + \bar\epsilon_{\max}$. Since $f(x^1) \le f(\bar x)$, we have $x^1, \bar x \in T_\alpha$ (cf. (2.1)). First, suppose $x^k \in T_\alpha$. Since $|x^{k+1} - x^k| \le \nu_k|g^k| \le \bar\rho$ by (3.3) and our assumption,

$$|x^{k+1} - \bar x| \le |x^k - \bar x| + |x^{k+1} - x^k| \le \operatorname{diam}(T_\alpha) + \bar\rho \quad \text{if} \quad x^k \in T_\alpha. \tag{7.25}$$

Next, suppose $x^k \notin T_\alpha$. Then $f(x^k) > \max_{\bar x + B_{\bar\rho}} f + \epsilon_k$, since $\epsilon_k \le \bar\epsilon_{\max}$. Thus for $x = \bar x$ in Lemma 7.8, we have $f(x^k) > \omega_x(\bar\rho) + \epsilon_k$ (cf. (7.15)), so (7.17) yields $\rho_k > \bar\rho$, and then (3.4) or, equivalently, the first inequality of (7.18) with $\nu_k|g^k| \le \bar\rho$ gives $|x^{k+1} - \bar x| \le |x^k - \bar x|$. Combining this with (7.25) yields $|x^k - \bar x| \le \operatorname{diam}(T_\alpha) + \bar\rho$ for all k, since $x^1, \bar x \in T_\alpha$. ∎

In the framework of Proposition 7.13, we may use the following bounding strategy that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$, $\bar\rho > 0$, and $\bar\epsilon_{\max} \ge 0$, pick a positive sequence $\nu_{\max}^l \to 0$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu_{\max}^l$ and errors $\epsilon_k \le \bar\epsilon_{\max}$; if $\nu_k|g^k| > \bar\rho$ for some k, then increase l by 1, restart the algorithm, etc. Since the number of restarts is finite by Theorem 6.4, this strategy ensures the boundedness of $\{x^k\}$ and $\{g^k\}$. A special case of this strategy consists of picking a sequence $\nu_k \to 0$ and resetting x^{k+1} to x^1 whenever $\nu_k|g^k| > \bar\rho$ (as in [Sho79, Thm. 2.4]).

Remark 7.14. Proposition 7.13 also fills a gap in the proof of [Sho79, Thm. 2.4].

7.4. Fejér-type stepsizes. We now highlight a property of the *quadratic* scalings (7.6)–(7.9) and (7.13)–(7.14) based on $|g^k|^2$ which distinguishes them from the *linear* scalings (7.10)–(7.12) that use $|g^k|$.

COROLLARY 7.15. *Suppose that* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$ *and the algorithm employs a locally bounded oracle and stepsizes* $\nu_k := \lambda_k \mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^{\infty} \lambda_k = \infty$, *and* μ_k *chosen as in* (7.6)–(7.9) *or* (7.13)–(7.14). *If* $\lambda := \overline{\lim}_{k\to\infty} \lambda_k$ *is finite, then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar\gamma\lambda + \epsilon$, *where* (*cf.* (7.1)) $\bar\gamma$ *is at most* $\frac{1}{2}G$ *for* μ_k *chosen via* (7.6), *and* $\frac{1}{2}G^2$ *for* (7.7), $\frac{1}{2}$ *for* (7.8) *and* (7.13)–(7.14), *and* $\frac{1}{2}G^2$ *for* (7.9). *Consequently, we have* $\inf_k f(x^k) \le f_* + \frac{1}{2}\bar\gamma\lambda + \epsilon$ *if* λ *is finite whenever* $\inf_k f(x^k) > -\infty$.

Proof. This follows from Theorem 7.1(i) and the discussion in Example 7.3. ◻

Remark 7.16.

(i) Corollary 7.15 says that for the quadratic scalings (7.6)–(7.9) and (7.13)–(7.14), the asymptotic objective accuracy can be controlled by choosing the stepsize value λ a priori. In contrast, the asymptotic accuracy for the linear scalings (7.10)–(7.12) depends on the value of $\inf_{x\in S} \max_{x+B_{\lambda/2}} f$ (cf. Thm 7.9), which may be hard to guess.

(ii) The following adaptive choice of λ_k meets the requirements of Corollary 7.15. Select $\lambda_{\min} \in (0,\infty)$, $\kappa \in (0,1)$, and $\lambda_1 \ge \lambda_{\min}$. For each k, letting $f_{\mathrm{rec}}^k := \min_{j=1}^k f(x^j)$, choose

$$\lambda_{k+1} \in \begin{cases} [\lambda_{\min}, \infty) & \text{if} \quad f(x^{k+1}) \le f_{\mathrm{rec}}^k - \lambda_{\min}, \\ [\lambda_{\min}, \max\{\lambda_{\min}, \kappa\lambda_k\}] & \text{if} \quad f(x^{k+1}) > f_{\mathrm{rec}}^k - \lambda_{\min}. \end{cases} \tag{7.26}$$

Clearly, either $f_{\mathrm{rec}}^k \downarrow -\infty$ (and hence $f_* = -\infty$) or $\lambda_k = \lambda_{\min}$ for all large k.

Our quadratic scalings are related to *Fejér* stepsizes that reduce the distance to the solution set S_*. The latter stem from the observation that for $x \in S_*$ and $\epsilon_k = 0$, the optimal stepsize ν_k that minimizes the right-hand side of the estimate (3.1) has the form $\nu_k = \lambda_k\mu_k$ with $\lambda_k = f(x^k) - f_*$ and $\mu_k = |g^k|^{-2}$. Such stepsizes are analyzed below.

THEOREM 7.17. *Suppose that* $f_* > -\infty$ *and the algorithm employs a locally bounded oracle and stepsizes*

$$\nu_k := \kappa_k \left[f(x^k) - f_* \right] |g^k|^{-2} \quad \textit{with} \quad \kappa_k \in [\kappa_{\min}, \kappa_{\max}] \subset (0,2). \tag{7.27}$$

(i) *If* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k$ *is finite, then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \frac{2}{2-\kappa_{\max}}\epsilon$.
(ii) *If the solution set* S_* *is nonempty and for all* k

$$\epsilon_k \le \tfrac{1}{2}\kappa_\epsilon(2-\kappa_k) \left[f(x^k) - f_* \right] \quad \textit{with} \quad \kappa_\epsilon \in [0,1), \tag{7.28}$$

then $\{x^k\}$ *converges to some solution* $x^\infty \in S_*$ *and* $\lim_{k\to\infty} f(x^k) = f_*$.

Proof. (i) For contradiction, suppose $\frac{2-\kappa_{\max}}{2} \underline{\lim}_{k\to\infty} \lambda_k > \epsilon$, where $\lambda_k := f(x^k) - f_*$. Since $\epsilon := \overline{\lim}_k \epsilon_k \ge 0$ and $f_* := \inf_S f$, there exist $\kappa \in (0,1)$, $x \in S$, and k_ϵ such that

$$\kappa \frac{2-\kappa_{\max}}{2} \lambda_k \ge f(x) - f_* + \epsilon_k \quad \forall k \ge k_\epsilon. \tag{7.29}$$

Using the fact that $\lambda_k := f(x^k) - f_* \ge 0$, (7.27), (7.29), and again (7.27) in (3.1) yields

$$
\begin{aligned}
|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k \left[\, f_* - f(x) - \epsilon_k + f(x^k) - f_* - \tfrac{1}{2}\nu_k |g^k|^2 \,\right] \\
&= -2\nu_k \left[\, f_* - f(x) - \epsilon_k + \lambda_k - \tfrac{1}{2}\kappa_k\lambda_k \,\right] \\
&\le -2\nu_k (1-\kappa)\tfrac{2-\kappa_{\max}}{2}\lambda_k \\
&\le -\kappa_{\min}(1-\kappa)(2-\kappa_{\max})\lambda_k^2/|g^k|^2 < 0 \qquad \forall k \ge k_\epsilon. \qquad (7.30)
\end{aligned}
$$

By (7.30), $\{x^k\}$ is bounded and $\sum_k \lambda_k^2/|g^k|^2 < \infty$. Hence $\sup_k |g^k| < \infty$ (because the oracle is locally bounded and $\epsilon < \infty$) and $\lambda_k^2/|g^k|^2 \to 0$ yields $\lambda_k \to 0$, a contradiction.

(ii) For any $x \in S_*$, using (7.28) and (7.27) as in (7.30) yields

$$
\begin{aligned}
|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k \left[\, \tfrac{1}{2}(2-\kappa_k)\lambda_k - \epsilon_k \,\right] \\
&\le -2\nu_k (1-\kappa_\epsilon)\tfrac{2-\kappa_{\max}}{2}\lambda_k \\
&\le -\kappa_{\min}(1-\kappa_\epsilon)(2-\kappa_{\max})\lambda_k^2/|g^k|^2 < 0 \qquad \forall k \ge 1, \qquad (7.31)
\end{aligned}
$$

so again $\{x^k\}$ is bounded and $\lambda_k/|g^k| \to 0$. Then $\epsilon := \overline{\lim}_k\, \epsilon_k < \infty$ by (7.28) (since f is continuous because the oracle is bounded), and as in (i) we get $\lambda_k := f(x^k) - f_* \to 0$. Further, $\{x^k\}$ has a cluster point $x^\infty \in S$ with $f(x^\infty) \le f_*$ (since S and f are closed), i.e., $x^\infty \in S_*$. Setting $x = x^\infty$ in (7.31) shows that $|x^k - x^\infty| \downarrow 0$, i.e., $x^k \to x^\infty$. $\square$

Remark 7.18. In contrast to standard results, Theorem 7.17(i) *does not* assume nonemptiness of the solution set S_*. Theorem 7.17(ii) subsumes [Pol69, Thm. 1] (where $\epsilon_k \equiv 0$) and [Brä95, Thm. 2.4] (for a special oracle). As in [Brä95, sect. 2], condition (7.28) may be replaced by $\inf_k \kappa_k(2-\kappa_k-2\epsilon_k/\lambda_k) > 0$ with (7.27) relaxed to $\kappa_k \in [0,2]$.

Since the optimal value f_* in (7.27) is usually unknown, it may be replaced by a *target level* $f_{\text{lev}}^k := f_{\text{rec}}^k - \tilde\delta_k$ with $\tilde\delta_k$ updated as in (7.26); the resulting scheme is analyzed below.

THEOREM 7.19. *Suppose that* $\epsilon := \overline{\lim}_{k\to\infty}\, \epsilon_k < \infty$ *and the algorithm employs a locally bounded oracle and stepsizes* $\nu_k := \lambda_k \mu_k$ *with*

$$\lambda_k := f(x^k) - f_{\text{lev}}^k, \quad f_{\text{lev}}^k := f_{\text{rec}}^k - \tilde\delta_k, \qquad (7.32)$$

$$\mu_k := \kappa_k |g^k|^{-2}, \quad \kappa_k \in [\kappa_{\min}, \kappa_{\max}], \qquad (7.33)$$

where $f_{\text{rec}}^k := \min_{j=1}^k f(x^j)$, $0 < \kappa_{\min} \le \kappa_{\max} \le 2$, *and* $\tilde\delta_k > 0$ *is such that* $\tilde\delta := \overline{\lim}_{k\to\infty}\, \tilde\delta_k \in (0,\infty)$ *whenever* $f_{\text{rec}}^\infty := \inf_k f(x^k) > -\infty$ (*e.g.*, $\tilde\delta_k \equiv \tilde\delta > 0$). *Then either* $f_{\text{rec}}^\infty = -\infty = f_*$ *or* $f_{\text{rec}}^\infty \le f_* + \tilde\delta + \epsilon$ *with* $\tilde\delta < \infty$.

Proof. If $f_{\text{rec}}^\infty = -\infty$, then $f_* \le \inf_k f(x^k) = -\infty$, so assuming $f_{\text{rec}}^\infty > -\infty$, suppose $f_{\text{rec}}^\infty > f_* + \epsilon + \tilde\delta$. Then there exist $x \in S$ and $v > 0$ such that $f_{\text{rec}}^k \ge f(x) + \epsilon + \tilde\delta + v$ for all k, so using (7.32) with $\tilde\delta := \overline{\lim}_k\, \tilde\delta_k$ and $\epsilon := \overline{\lim}_k\, \epsilon_k$ we deduce the existence of k_v such that

$$f_{\text{lev}}^k - f(x) - \epsilon_k = f_{\text{rec}}^k - f(x) - \tilde\delta_k - \epsilon_k \ge \tilde\delta - \tilde\delta_k + \epsilon - \epsilon_k + v \ge \tfrac{1}{2}v \qquad (7.34)$$

for all $k \ge k_v$. Since $\lambda_k \ge \tilde\delta_k > 0$ by (7.32) and $\mu_k|g^k|^2 \le \kappa_{\max}$ by (7.33), we have $\nu_k|g^k|^2 \le \kappa_{\max}\lambda_k$. Hence using (7.32), (7.34), and $\kappa_{\max} \le 2$ in the estimate (3.1) yields

$$
\begin{aligned}
|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k\left[\, f_{\rm lev}^k - f(x) - \epsilon_k + f(x^k) - f_{\rm lev}^k - \tfrac12\nu_k|g^k|^2 \,\right]\\
&\le -2\nu_k\left[\, f_{\rm lev}^k - f(x) - \epsilon_k + \lambda_k - \tfrac12\kappa_{\max}\lambda_k \,\right]\\
&\le -\nu_k\left[\, v + (2-\kappa_{\max})\lambda_k \,\right] \le -v\nu_k < 0 \qquad \forall k \ge k_v. \qquad (7.35)
\end{aligned}
$$

By (7.35), $\{x^k\}$ is bounded and $\sum_k \nu_k < \infty$. Hence $\hat G := \sup_k |g^k| < \infty$ (because the oracle is locally bounded and $\epsilon < \infty$) and $\lim_k \nu_k = 0$. However, $\nu_k := \lambda_k\mu_k \ge \tilde\delta_k\kappa_{\min}\hat G^{-2}$ by (7.32)–(7.33), where $\kappa_{\min} > 0$, so we get $\tilde\delta := \overline{\lim}_k\, \tilde\delta_k = 0$, a contradiction. □

Remark 7.20.

(i) The following adaptive choice of $\tilde\delta_k$ meets the requirements of Theorem 7.19. Select $\tilde\delta_{\min} \in (0,\infty)$, $\kappa \in (0,1)$, and $\tilde\delta_1 \ge \tilde\delta_{\min}$. For each k, choose

$$
\tilde\delta_{k+1} \in \begin{cases} [\tilde\delta_{\min}, \infty) & \text{if} \quad f(x^{k+1}) \le f_{\rm rec}^k - \tilde\delta_{\min},\\ \left[\tilde\delta_{\min}, \max\left\{\tilde\delta_{\min}, \kappa\tilde\delta_k\right\}\right] & \text{if} \quad f(x^{k+1}) > f_{\rm rec}^k - \tilde\delta_{\min}. \end{cases} \qquad (7.36)
$$

Clearly, either $f_{\rm rec}^k \downarrow -\infty$ (and hence $f_* = -\infty$) or $\tilde\delta_k = \tilde\delta_{\min}$ for all large k.

(ii) A special case of (7.36) introduced in [NeB01, eq. (2.19)] is to set $\tilde\delta_{k+1} := \eta\tilde\delta_k$ if $f(x^{k+1}) \le f_{\rm lev}^k$, $\tilde\delta_{k+1} := \max\{\tilde\delta_{\min}, \kappa\tilde\delta_k\}$ otherwise, where $\eta \in [1,\infty)$. For this case Theorem 7.19 subsumes [NeB01, Rem. 2.1] (where $\epsilon_k \equiv 0$ and $\kappa_{\max} < 2$ in (7.33)). In the exact case ($\epsilon_k \equiv 0$) similar schemes with nonvanishing level gaps are considered in [Kiw96b, Thm. 4.4], [Kiw98, Thm. 4.2], and [SCT00]; vanishing level gaps are studied in [Brä93, GoK99, KLL99b, NeB01].

8. Efficiency estimates. In order to derive efficiency estimates, in this section we assume that the optimal set S_* is nonempty and that the sequences $\{x^k\}$, $\{g^k\}$, and $\{\epsilon_k\}$ are bounded.

For some stepsizes, sharper estimates may be derived by replacing the index $j = 1$ in (3.2), (3.7), (3.8), and (3.10) by $j = k'$, where k' depends on k, e.g., $k' := \lceil \frac12 k\rceil$. Thus for

$$
\bar f_k := \sum_{j=k'}^k \nu_j f(x^j)/\nu_{\rm sum}^k,\ \bar x^k := \sum_{j=k'}^k \nu_j x^j/\nu_{\rm sum}^k,\ \bar\epsilon_k := \sum_{j=k'}^k \nu_j\epsilon_j/\nu_{\rm sum}^k,\ \nu_{\rm sum}^k := \sum_{j=k'}^k \nu_j, \qquad (8.1)
$$

replacing 1 by k' in (3.2) and using $x := P_{S_*}x^{k'}$ yields the estimate

$$
\bar f_k - f_* \le \Delta_k + \bar\epsilon_k \quad \text{with} \quad \Delta_k := \frac{d_{S_*}^2(x^{k'}) + \sum_{j=k'}^k \nu_j^2|g^j|^2}{2\sum_{j=k'}^k \nu_j}. \qquad (8.2)
$$

This is indeed an *accuracy estimate*, since we still have (cf. (3.9), (3.14))

$$
f(\bar x^k) \le \bar f_k \quad \text{and} \quad \min\left\{\, f(x^j) : j = k' : k \,\right\} \le \bar f_k. \qquad (8.3)
$$

Our efficiency estimates involve the (problem and algorithm-dependent) quantities

$$
\hat D := \sup\nolimits_k d_{S_*}(x^k) \quad \text{and} \quad \hat G := \sup\nolimits_k |g^k|. \qquad (8.4)
$$

To provide freedom for implementations, we allow for additional scaling factors

$$D_k \in [\,D_{\min}, D_{\max}\,] \subset (0,\infty) \quad \text{and} \quad G_k \in [\,G_{\min}, G_{\max}\,] \subset (0,\infty). \tag{8.5}$$

For a fixed $s \in [1/2, 1]$, we consider the following stepsizes and their *efficiency factors*:

$$\nu_k := \frac{D_k k^{-s}}{\max\{|g^k|, |g^k|^2/G_k\}} \text{ with } c_{(8.6)} := \max\{\hat{G}, G_{\min}, \hat{G}^2/G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.6}$$

$$\nu_k := \frac{D_k k^{-s}}{\max\{G_k, |g^k|^2/G_k\}} \text{ with } c_{(8.7)} := \max\{G_{\max}, \hat{G}^2/G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.7}$$

$$\nu_k := \frac{D_k k^{-s}}{|g^k|} \text{ with } c_{(8.8)} := \max\{\hat{G}, G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.8}$$

$$\nu_k := \frac{D_k k^{-s}}{G_k} \text{ with } c_{(8.9)} := G_{\max}\frac{\hat{D}^2 + D_{\max}^2(\hat{G}/G_{\min})^2}{D_{\min}}, \tag{8.9}$$

where $|g^k|$ is replaced by $G_{\min}$ if $|g^k| = 0$. For such stepsizes, the sums involved in (8.2) may be bounded via the following lemma.

LEMMA 8.1. *For $k \ge 1$ and $s \in [1/2, 1]$, we have the following statements:*

(i) $\sum_{j=\lceil \frac{1}{2}k \rceil}^{k} j^{-2s} \le 1 + \ln 2$ *and* $\sum_{j=\lceil \frac{1}{2}k \rceil}^{k} j^{-s} \ge (2 - 2^{1/2})(k+1)^{1-s}$.

(ii) $\sum_{j=1}^{k} j^{-2s} \le \min\{\frac{2s}{2s-1}, 1 + \ln k\}$ *and* $\sum_{j=1}^{k} j^{-s} \ge \max\{\ln(k+1), (2 - 2^{1/2})(k+1)^{1-s}\}$.

Proof. For $s \in (1/2, 1)$, this follows from standard integration arguments (cf. [Nes89, p. 157]), using the facts that $\frac{2^{s-1}-1}{s-1} \ge 2 - 2^{1/2}$ for (i), $\frac{k^{1-2s}-1}{1-2s} \le \ln k$, and $\frac{(k+1)^{1-s}-1}{1-s} \ge \ln(k+1)$ for (ii); the rest follows by continuity. ∎

We may now state our efficiency estimates for the stepsizes (8.6)–(8.9).

THEOREM 8.2. *For a fixed $s \in [1/2, 1]$, consider any stepsize rule from (8.6)–(8.9) and its efficiency factor c (e.g., $c := c_{(8.6)}$ for (8.6)). Then for each k we have*

$$\bar{f}_k - f_* \le \bar{\epsilon}_k + \begin{cases} \dfrac{(1+\ln 2)c}{(4 - 2^{3/2})(k+1)^{1-s}} & \text{if} \quad k' = \left\lceil \dfrac{1}{2}k \right\rceil, \\[2ex] \dfrac{\min\left\{\frac{2s}{2s-1}, 1 + \ln k\right\} c}{\max\left\{2\ln(k+1), (4 - 2^{3/2})(k+1)^{1-s}\right\}} & \text{if} \quad k' = 1. \end{cases} \tag{8.10}$$

If the errors satisfy $\epsilon_k \le C_\epsilon k^{-s}$ for some constant C_ϵ, and the stepsizes are chosen via (8.7) or (8.9), then we also have

$$\bar{\epsilon}_k \le \begin{cases} \dfrac{(1+\ln 2)C_\epsilon c_\epsilon}{(2 - 2^{1/2})(k+1)^{1-s}} & \text{if} \quad k' = \left\lceil \dfrac{1}{2}k \right\rceil, \\[2ex] \dfrac{\min\left\{\frac{2s}{2s-1}, 1 + \ln k\right\} C_\epsilon c_\epsilon}{\max\left\{\ln(k+1), (2 - 2^{1/2})(k+1)^{1-s}\right\}} & \text{if} \quad k' = 1, \end{cases} \tag{8.11}$$

where $c_\epsilon := \frac{\max\{G_{\max}, \hat{G}^2/G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for (8.7) and $c_\epsilon := \frac{D_{\max}G_{\max}}{D_{\min}G_{\min}}$ for (8.9); also (8.11) holds with $c_\epsilon := \frac{\max\{G_{\min}, \hat{G}^2/G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for (8.6) and $c_\epsilon := \frac{\max\{\hat{G}, G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for (8.8) provided that $|g^k|$ is replaced by $\max\{|g^k|, G_{\min}\}$ in the stepsizes of (8.6) and (8.8), in which case the bound (8.10) remains valid.

Proof. For (8.10), it suffices to bound Δ_k in (8.2) by using $d_{S_*}(x^{k'}) \le \hat{D}$ (cf. (8.4)), and then $|g^k| \le \hat{G}$ and (8.5) together with Lemma 8.1 for the sums. For (8.11), the sums of $\bar{\epsilon}_k$ (cf. (8.1)) are estimated in a similar way. ∎

The estimates (8.10) and (8.11) combine nicely into an *overall* efficiency estimate.

Remark 8.3.

(i) It follows from general complexity results [BTMN01, Prop. 4.1] that for $\epsilon_k \equiv 0$ and n large enough, a *lower* bound on $\min_{j=1}^k f(x^j) - f_*$ is of order $O(k^{-1/2})$. Since (8.3) and (8.10) imply an *upper* bound of the same order for $s = 1/2$ and $k' = \lceil \frac{1}{2}k \rceil$, this choice is *optimal* from the complexity viewpoint. The switch from $k' = \lceil \frac{1}{2}k \rceil$ to $k' = 1$ degrades the bound moderately to $O(k^{-1/2} \ln k)$, but the popular choice of $s = 1$ has a much worse bound of $O(1/\ln k)$. On the other hand, for $s = 1/2$ we cannot have $\sum_k \nu_k^2 < \infty$ as required for convergence of $\{x^k\}$ in Theorem 3.4; however, choosing s slightly larger than $1/2$ combines the best of both worlds: convergence of $\{x^k\}$ and efficiency of order $O(k^{s-1})$ comparable to $O(k^{-1/2})$.

(ii) The stepsize (8.6) corresponds to (7.6) (with $\lambda_k := D_k k^{-s}$), (8.7) corresponds to both (7.7) and (7.8) (with $\lambda_k := (D_k/G_k)k^{-s}$ and $\lambda_k := D_k G_k k^{-s}$, respectively), and (8.8) corresponds to (7.10). For these stepsizes Theorems 7.1 and 7.11 ensure finiteness of $\hat{D}$ and $\hat{G}$ in (8.4) under reasonable conditions. The stepsize (8.9) may need the bounding strategies of section 6, e.g., for picking $D_{\max}$ small enough.

(iii) The efficiency factors of (8.6)–(8.9) are of order $2\hat{G}\hat{D}$ when $D_{\min} \approx D_{\max} \approx \hat{D}$, $G_{\min} \approx G_{\max} \approx \hat{G}$, but in general the values of $\hat{D}$ and $\hat{G}$ in (8.4) are stepsize-dependent.

In the language of Theorem 7.1(i), nonvanishing stepsizes ensure only asymptotic objective accuracy of order $\tilde{\delta} \approx \bar{\gamma}\lambda$ (for ϵ_k sufficiently small). In this context, efficiency is understood in terms of bounds on the relative accuracy $(\Delta_k - \tilde{\delta})/\tilde{\delta}$ (cf. (8.2)–(8.3)). Roughly speaking, for reasonable stepsizes such bounds have the form $(\hat{\Delta}/2\tilde{\delta})^2/k$, where $\hat{\Delta}$ measures the variation of f; a more precise statement is given below.

PROPOSITION 8.4. *For fixed $\lambda > 0$, $G > 0$, $D := d_{S_*}(x^1)$, and $\hat{G} := \sup_k |g^k|$, the stepsizes ν_k exhibited below have the following given efficiency bounds on Δ_k (cf.* (8.2)–(8.3) *with $k' = 1$):*

$$\nu_k := \frac{\lambda}{\max\{|g^k|, |g^k|^2/G\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}G\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{(G\lambda)^2 k}\right), \tag{8.12}$$

$$\nu_k := \frac{\lambda}{\max\{1, |g^k|^2/G^2\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}G^2\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{(G^2\lambda)^2 k}\right), \tag{8.13}$$

$$\nu_k := \frac{\lambda}{\max\{G^2, |g^k|^2\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{\lambda^2 k}\right), \tag{8.14}$$

$$\nu_k := \frac{\lambda}{|g^k|^2} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\lambda\left(1 + \frac{\hat{G}^2 D^2}{\lambda^2 k}\right), \tag{8.15}$$

$$\nu_k := \frac{\lambda}{|g^k|} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\hat{G}\lambda\left(1 + \frac{\hat{G}^2 D^2}{(\hat{G}\lambda)^2 k}\right), \tag{8.16}$$

$$\nu_k := \lambda \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\hat{G}^2\lambda\left(1 + \frac{\hat{G}^2 D^2}{(\hat{G}^2\lambda)^2 k}\right). \tag{8.17}$$

Here we assume that $|g^k|$ is replaced by G in (8.12) *and* (8.15)–(8.16) *whenever $|g^k| = 0$ and for* (8.15)–(8.16) *that G is reset to $|g^k|$ when $|g^k|$ becomes nonzero.*

Proof. Recalling the definition (8.2) of Δ_k, simple calculations yield the conclusion. $\square$

9. Analysis of the incremental subgradient method.

9.1. Basic incremental estimates. Throughout this section, $\{x^k\}$, $\{\nu_k\}$, $\{x_i^k\}$, $\{\epsilon_i^k\}$, and $\{g_i^k\}$ denote the sequences involved in the incremental subgradient iteration (1.4). Further, for each k, we let

$$f_{\rm inc}^k := \sum_{i=1}^m f_i(x_i^k), \tag{9.1}$$

$$\epsilon_k := \sum_{i=1}^m \epsilon_i^k, \tag{9.2}$$

$$\bar{C}_k := \sum_{i=1}^m \bar{C}_{ik} \quad \text{with} \quad \bar{C}_{ik} := \max\left\{ |g_i^k|, |\bar{g}_i^k| \right\} \quad \text{for some} \quad \bar{g}_i^k \in \partial f_i^S(x^k). \tag{9.3}$$

Note that the *incremental* objective value $f_{\rm inc}^k$ is a natural estimate for $f(x^k)$, and the additional subgradients $\bar{g}_i^k$ provide only bounds on $f(x^k) - f_{\rm inc}^k$ (cf. (9.8), (9.11)).

We start by extending the basic estimates of Lemma 3.1 to the incremental case.

LEMMA 9.1. *For each x and $k \geq 1$, we have*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \leq -2\nu_k \left[f(x^k) - f_S(x) - \epsilon_k - \tfrac{1}{2}\bar{C}_k^2 \nu_k \right], \tag{9.4}$$

$$\frac{\sum_{j=1}^k \nu_j f(x^j)}{\sum_{j=1}^k \nu_j} - f_S(x) \leq \frac{\frac{1}{2}|x^1 - x|^2 + \sum_{j=1}^k \frac{1}{2}\nu_j^2 \bar{C}_j^2 + \sum_{j=1}^k \nu_j \epsilon_j}{\sum_{j=1}^k \nu_j}, \tag{9.5}$$

$$|x^{k+1} - x^k| \leq \nu_k \bar{C}_k, \tag{9.6}$$

$$|x^{k+1} - x|^2 - |x^k - x|^2 \leq -2\nu_k \left[f_{\rm inc}^k - f_S(x) - \epsilon_k - \tfrac{1}{2}\nu_k \sum_{i=1}^m |g_i^k|^2 \right], \tag{9.7}$$

$$f(x^k) - f_{\rm inc}^k \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} \bar{C}_{jk}, \tag{9.8}$$

$$f_{\rm inc}^k - f(x^k) - \epsilon_k \leq \nu_k \sum_{i=1}^m |g_i^k| \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} \bar{C}_{jk}, \tag{9.9}$$

$$|x_i^k - x^k| \leq \nu_k \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{j=1}^{i-1} \bar{C}_{jk} \quad \textit{for } i = 1\colon m+1. \tag{9.10}$$

Proof. Let $x \in S$, $r_{ik} := |x_i^k - x|$. Using the nonexpansiveness of P_S and (1.4) gives

$$\begin{aligned} r_{i+1,k}^2 &\le |x_i^k - \nu_k g_i^k - x|^2 = r_{ik}^2 - 2\nu_k \left\langle g_i^k, x_i^k - x \right\rangle + \nu_k^2 |g_i^k|^2 \\ &\le r_{ik}^2 + 2\nu_k \left[f_i(x) - f_i(x_i^k) + \epsilon_i^k \right] + \nu_k^2 |g_i^k|^2; \end{aligned}$$

sum up and use $r_k := |x^k - x|$, $x^{k+1} := x_{m+1}^k$, and (9.1)–(9.2) to get (9.7). Since $|x_{i+1}^k - x^k| \le |x_i^k - x^k| + |x_{i+1}^k - x_i^k|$, where $|x_{i+1}^k - x_i^k| \le \nu_k |g_i^k|$ by (1.4), (9.10) follows by induction. Summing $f_i(x^k) - f_i(x_i^k) \le \left\langle \bar{g}_i^k, x^k - x_i^k \right\rangle$ (cf. (9.3)) and using (9.1) and (9.10), we obtain

$$f(x^k) - f_{\mathrm{inc}}^k = \sum_i \left[f_i(x^k) - f_i(x_i^k) \right] \le \sum_i |\bar{g}_i^k| |x_i^k - x^k| \le \nu_k \sum_i \bar{C}_{ik} \sum_{j<i} |g_j^k| \tag{9.11}$$

and hence (9.8); similarly, summing $f_i(x_i^k) - f_i(x^k) - \epsilon_i^k \le \left\langle g_i^k, x_i^k - x^k \right\rangle$ (cf. (1.4)) gives (9.9). Then (9.7), (9.8), and (9.3) yield (9.4), since $2\sum_i \bar{C}_{ik} \sum_{j<i} \bar{C}_{jk} + \sum_i \bar{C}_{ik}^2 = \bar{C}_k^2$. Summing up (9.4) gives (9.5). For $f_S(x) = \infty$, (9.4), (9.5), and (9.7) are trivial. Finally, (9.6) follows from (9.10) with $i = m+1$, using $x^{k+1} := x_{m+1}^k$ and (9.3). ∎

9.2. General incremental convergence results. All the convergence results of sections 3 and 4 extend easily to the incremental method.

COROLLARY 9.2. *Theorems* 3.2, 3.4, 3.6, 4.1, *and Corollary* 4.2 *hold for the incremental subgradient method* (1.4) *with* $|g^k|$ *replaced by* $\bar{C}_k$ (*so that* $\gamma_k := \frac{1}{2}\bar{C}_k^2 \nu_k$ *in* (3.5) *and* $C := \overline{\lim}_{k\to\infty} \bar{C}_k$ *in Theorems* 3.2(vi) *and* 4.1(iv)).

Proof. Comparing (3.1)–(3.3) with (9.4)–(9.6), we may replace $|g^k|$ by $\bar{C}_k$ in the proofs of sections 3.2–3.3 and section 4. ∎

We now give a more refined version of Corollary 4.2 for the incremental case that employs a slightly weaker assumption (boundedness of $|g_i^k|$ instead of $\max\{|g_i^k|, |\bar{g}_i^k|\}$).

LEMMA 9.3. *Suppose that* f_S *is coercive,* $\hat{\nu} := \sup_k \nu_k < \infty$, $\hat{\epsilon} := \sup_k \epsilon_k < \infty$, *and* $C_i := \sup_k |g_i^k| < \infty$ *for all* i. *Then* $\{x^k\}$ *and* $\{x_i^k\}$ *are bounded for all* i.

Proof. Let $x \in S_*$, $C := \sum_i C_i$, $\sigma := C\hat{\nu}$, and $\alpha := f_* + \hat{\epsilon} + \frac{1}{2}C^2\hat{\nu}$. Since $f(x) = f_*$ and f_S is coercive, x lies in the bounded set $T_{\alpha,\sigma}$ (cf. (2.3)). First, suppose that $f_{\mathrm{inc}}^k \le \alpha$. By (9.10) with $\nu_k \le \hat{\nu}$, we have $\max_i |x_i^k - x^k| \le \nu_k C \le \sigma$. Hence $x^k \in T_{\alpha,\sigma}$ (cf. (2.3) and (9.1)) and $|x^{k+1} - x^k| \le \sigma$ (since $x^{k+1} := x_{m+1}^k$). Thus

$$|x^{k+1} - x| \le |x^k - x| + |x^{k+1} - x^k| \le \mathrm{diam}(T_{\alpha,\sigma}) + \sigma \quad \text{if} \quad f_{\mathrm{inc}}^k \le \alpha. \tag{9.12}$$

Second, if $f_{\mathrm{inc}}^k > \alpha$, i.e., $f_{\mathrm{inc}}^k > f(x) + \hat{\epsilon} + \frac{1}{2}C^2\hat{\nu}$, then by using the bounds $\nu_k \le \hat{\nu}$, $\epsilon_k \le \hat{\epsilon}$, and $\sum_i |g_i^k|^2 \le \sum_i C_i^2 \le C^2$ in (9.7), we obtain

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[\tfrac{1}{2}C^2\hat{\nu} + \hat{\epsilon} - \epsilon_k - \tfrac{1}{2}\nu_k C^2 \right] \le 0 \quad \text{if} \quad f_{\mathrm{inc}}^k > \alpha. \tag{9.13}$$

Combining (9.12) and (9.13) gives $|x^k - x| \le \max\{\mathrm{diam}(T_{\alpha,\sigma}) + \sigma, |x^1 - x|\}$ for all k. Thus $\{x^k\}$ is bounded, and so are $\{x_i^k\}$ for all i, since $\max_i |x_i^k - x^k| \le \sigma$. ∎

Of course, in the incremental case Definition 6.1 is replaced by the following definition.

Definition 9.4. *We say that the algorithm employs a* locally bounded oracle *if $g_i^k = g_i(x^k, \epsilon_i^k)$ and $\bar{g}_i^k = g_i(x^k, 0)$ for all i and k, where the mappings $S \times \mathbb{R}_+ \ni (x, \epsilon) \mapsto g_i(x, \epsilon) \in \partial_\epsilon f_i^S(x)$ are locally bounded.*

The following result complements Lemma 9.3 and enables us to extend Theorem 5.1 to the incremental method.

Lemma 9.5. *Suppose that $\{x^k\}$ is bounded and $\hat{\nu} := \sup_k \nu_k < \infty$. Then we have the following statements:*

(i) *If the oracle is locally bounded and $\hat{\epsilon} := \sup_k \epsilon_k < \infty$, then $\{x_i^k\}$ is bounded for all i, and $\sup_k \bar{C}_k < \infty$.*

(ii) *If $\sup_k \bar{C}_k < \infty$, then $\{x_i^k\}$ is bounded for all i.*

Proof. (i) By Definition 9.4, $\{\bar{g}_i^k = g_i(x^k, 0)\}$ is bounded for all i. Assuming $C_j := \sup_k \bar{C}_{jk} < \infty$ for $j < i$, by (9.10) we have $|x_i^k - x^k| \le \hat{\nu} \sum_{j<i} C_j$ ($x_i^k = x^k$ if $i = 1$). Thus $\{x_i^k\}$ is bounded, and so is $\{g_i^k = g_i(x_i^k, \epsilon_i^k)\}$ because the oracle is locally bounded. Hence, by (9.3), $C_i := \sup_k \bar{C}_{ik}$ is finite. The rest follows by induction, with $\sup_k \bar{C}_k \le \sum_i C_i$.

(ii) This follows from (9.3) and (9.10) with $\nu_k \le \hat{\nu}$. ∎

Corollary 9.6. *Theorem* 5.1 *holds for the incremental subgradient method* (1.4) *with $|g^k|$ replaced by $\bar{C}_k$ (so that $\gamma_k := \frac{1}{2}\bar{C}_k^2 \nu_k$) and $\underline{R}$ redefined as $\underline{R} := \sup_{i,k} |x_i^k|$.*

Proof. The assumptions of Theorem 5.1 and Lemma 9.5 yield $\underline{R} < \infty$. Next, in the proof of Theorem 5.1, we may replace S and f_i^S in (1.4) by $S' := S \cap B_R$ and $f_i^{S'} := f_i^S + \mathrm{I}_{\mathrm{B_R}}$, since $\{x_i^k\} \subset S'$, whereas $g_i^k \in \partial_{\epsilon_i^k} f_i^S(x_i^k)$ implies $g_i^k \in \partial_{\epsilon_i^k} f_i^{S'}(x_i^k)$. In view of Corollary 9.2, the proof may be finished as before. ∎

Theorems 7.17 and 7.19 also may be extended to the incremental case.

Corollary 9.7. *Theorems* 7.17 *and* 7.19 *hold for the incremental subgradient method* (1.4) *if $|g^k|$ in* (7.27) *and* (7.33) *is replaced by a constant $C \in (0, \infty)$ such that $C \ge \sup_k \bar{C}_k$.*

Proof. Replace $|g^k|$ by C in the original proofs, invoking (9.4) instead of (3.1). ∎

Remark 9.8. Our framework is more general than that of [NeB01, sect. 2], where each f_i is finite-valued and $g_i^k \in \partial f_i(x_i^k)$ in (1.4); i.e., $\epsilon_i^k \equiv 0$ and the oracle is locally bounded. The basic assumption of [NeB01, Ass. 2.1] is $\sup_k \bar{C}_i^k < \infty$ for all i. Theorem 3.2(ii), (vi) subsumes [NeB01, Props. 2.1–2.2] (with $C := \sup_k \bar{C}_k$), Theorem 3.4 subsumes [NeB01, Prop. 2.4], and Theorem 4.1(ii) subsumes [NeB01, Prop. 2.3] (with $\nu = 0$). Corollary 9.7 subsumes [NeB01, Props. 2.5–2.6].

9.3. Incremental bounding strategies. We now extend Theorems 6.3 and 6.4 to the incremental case.

Theorem 9.9. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Fix any point $\bar{x} \in S$ and a tolerance $\bar{\delta} \in (0, \infty)$. Then there exist thresholds $\bar{\nu}_{\max} > 0$ and $\bar{\epsilon}_{\max} > 0$ with the following property: If the algorithm starts from a point $x^1 \in T_{f(\bar{x})}$ (e.g., $x^1 = \bar{x}$) and employs stepsizes $\nu_k \le \bar{\nu}_{\max}$ and errors $\epsilon_k \le \bar{\epsilon}_{\max}$ for all k, then x^k stays in the bounded trench $T_{f(\bar{x})+\bar{\delta}}$ and $f_{\mathrm{inc}}^k \le f(\bar{x}) + 2\bar{\delta}$ for all k, and there exist $C_i < \infty$ such that $\bar{C}_{ik} := \max\{|g_i^k|, |\bar{g}_i^k|\} \le C_i$ and $|x_i^k - x^k| \le \nu_k \sum_{j<i} C_j$ for all k and i.*

Proof. Let $\beta := f(\bar{x})$, $\bar{\alpha} := \beta + \bar{\delta}$. Since the oracle is locally bounded, f_S is continuous on S (cf. Remark 6.2(iii)). By Lemma 2.4(ii), there exists $\bar{\rho} > 0$ such that $S \cap (T_\beta + B_{3\bar{\rho}}) \subset T_{\bar{\alpha}}$, whereas by Lemma 2.4(i) there is $\alpha \in (\beta, \bar{\alpha})$ such that

$T^{\alpha}_{\beta} \subset T_\beta + B_{\bar\rho}$; thus

$$S \cap \left(T^{\alpha}_{\beta} + B_{\bar\rho}\right) \subset S \cap (\,T_\beta + B_{2\bar\rho}\,) \subset S \cap (\,T_\beta + B_{3\bar\rho}\,) \subset T_{\bar\alpha}. \tag{9.14}$$

Let

$$\bar\epsilon_{\max} := \tfrac{1}{2}(\alpha - \beta), \tag{9.15}$$

$$C := \sum_i C_i \quad \text{with} \quad C_i := \sup\left\{\, |g_i(x,\epsilon)| : x \in S \cap (T_\beta + B_{3\bar\rho}), \epsilon \le \bar\epsilon_{\max} \,\right\}, \tag{9.16}$$

$$\bar\nu_{\max} := \min\left\{\, \bar\rho/C, (\alpha-\beta)/C^2 \,\right\}. \tag{9.17}$$

Note that $C < \infty$, since T_β is bounded and $\bar\epsilon_{\max} < \infty$.

Since $\{x^k\} \subset S$ and $f(x^1) \le f(\bar x) =: \beta$, we have $x^1 \in S \cap (T_\beta + B_{2\bar\rho})$.

Assuming $x^k \in S \cap (T_\beta + B_{2\bar\rho})$ for some $k \ge 1$, we now show that $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$. First, note that, by induction as for (9.10), we have $|g^k_i| \le C_i$ for $i = 1\colon m$ and

$$|x^k_i - x^k| \le \nu_k \sum_{j<i} |g^k_j| \le \bar\nu_{\max} \sum_{j<i} C_j \le \bar\rho \quad \text{for} \quad i = 1\colon m+1. \tag{9.18}$$

Indeed, suppose (9.18) holds for some $i \le m$. (Recall that $x^k_1 = x^k$.) Then $|x^k_{i+1} - x^k| \le |x^k_i - x^k| + |x^k_{i+1} - x^k_i|$, where $|x^k_{i+1} - x^k_i| \le \nu_k |g^k_i|$ by (1.4) with $|g^k_i| = |g_i(x^k_i, \epsilon^k_i)| \le C_i$ (cf. (9.16)) because $\epsilon^k_i \le \bar\epsilon_{\max}$ and $x^k_i \in T_\beta + B_{3\bar\rho}$ from $x^k \in T_\beta + B_{2\bar\rho}$ and $|x^k_i - x^k| \le \bar\rho$. Thus (9.18) holds for i increased by 1, with the final inequality due to (9.17). Further, (9.3) and (9.16) give $\bar C_{ik} \le C_i$ and $\bar C_k \le C$, using $|\bar g^k_i| = |g_i(x^k, 0)| \le C_i$. If $x^k \in T_\alpha$, then $T_\alpha \subset T^{\alpha}_{\beta}$ (cf. (2.2)), and the first inclusion of (9.14) and (9.18) with $x^{k+1} := x^k_{m+1}$ yield

$$x^{k+1} \in S \cap \left(\, x^k + B_{\bar\rho}\,\right) \subset S \cap (\,T_\alpha + B_{\bar\rho}\,) \subset S \cap \left(\,T^{\alpha}_{\beta} + B_{\bar\rho}\,\right) \subset S \cap (\,T_\beta + B_{2\bar\rho}\,).$$

Next, suppose $x^k \notin T_\alpha$, i.e.,

$$f(x^k) > \alpha. \tag{9.19}$$

Since $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, we have $|x^k - x| \le 2\bar\rho$ for $x = P_{T_\beta} x^k$. By (9.15) and (9.17),

$$\epsilon_k \le \bar\epsilon_{\max} \le \tfrac{1}{2}(\alpha-\beta) \quad \text{and} \quad \tfrac{1}{2}\bar C_k^2 \nu_k \le \tfrac{1}{2} C^2 \bar\nu_{\max} \le \tfrac{1}{2}(\alpha - \beta). \tag{9.20}$$

Using the estimate (9.4) with $f_S(x) \le \beta$ and the bounds (9.19) and (9.20), we obtain

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[\, f(x^k) - f(x) - \epsilon_k - \tfrac{1}{2}\bar C_k^2 \nu_k \,\right] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le 2\bar\rho$ with $x \in T_\beta$, so $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$.

Therefore, by induction, we have $x^k \in S \cap (T_\beta + B_{2\bar\rho}) \subset T_{\bar\alpha}$ (cf. (9.14)), $\bar C_{ik} \le C_i$, and (9.18) for all k. Finally, using (9.9) with $f(x^k) \le \bar\alpha$ and $\sum_i \bar C_{ik} \sum_{j<i} \bar C_{jk} \le \frac{1}{2}\bar C_k^2$ together with (9.20) gives $f^k_{\mathrm{inc}} \le \bar\alpha + \alpha - \beta \le \beta + 2\bar\delta$, since $\alpha < \bar\alpha := \beta + \bar\delta$. ∎

THEOREM 9.10. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Then for each $\beta \in (f_*, \infty)$ and $\bar\epsilon_{\max} \in [0, \infty)$ there exists $\bar\nu_{\max} > 0$*

such that if $f_S(x^1) \le \beta$*,* $\nu_k \le \bar\nu_{\max}$*, and* $\epsilon_k \le \bar\epsilon_{\max}$ *for all* k*, then* $\{x_i^k\}$*,* $\{g_i^k\}$ *and* $\{\bar g_i^k\}$ *are bounded for all* i*.*

Proof. Modify the proof of Theorem 9.9 as in the proof of Theorem 6.4. ◻

In view of Theorems 9.9–9.10, for the incremental method we may use the bounding strategy with the resetting test (6.7) or the strategy inspired by Theorem 6.4 with the test (6.8) replaced by $\max_i |x_i^k| > R_l$.

Yet another bounding strategy stems from the following result.

LEMMA 9.11. *Suppose that* f_S *is coercive and there exist* $\alpha \in \mathbb{R}$ *and* $\sigma \in \mathbb{R}_+$ *such that* $f_{\rm inc}^k \le \alpha$ *and* $\max_i |x_i^k - x^k| \le \sigma$ *for all* k*. Then* $\{x^k\}$ *is bounded.*

Proof. By (2.3) and (9.1), $\{x^k\}$ lies in the bounded set $T_{\alpha,\sigma}$ (cf. Lemma 2.5). ◻

Lemma 9.11 suggests the following bounding strategy with resets indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$, $\bar\delta \in (0,\infty)$, and $\bar\sigma \in (0,\infty)$, pick positive sequences $\nu_{\max}^l \to 0$ and $\epsilon_{\max}^l \to 0$ as $l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu_{\max}^l$ and errors $\epsilon_k \le \epsilon_{\max}^l$; if for some k

$$f_{\rm inc}^k > f(\bar x) + 2\bar\delta \quad \text{or} \quad \max_i |x_i^k - x^k| > \bar\sigma, \tag{9.21}$$

then increase l by 1, restart the algorithm, etc. Under the assumptions of Theorem 9.9, only finitely many resets occur, so Lemmas 9.5(i) and 9.11 imply the boundedness of $\{x_i^k\}$ and $\{\bar C_k\}$. (A special case of this strategy consists of using sequences $\nu_k \to 0$ and $\epsilon_k \to 0$, and resetting x^{k+1} to x^1 whenever (9.21) holds.)

9.4. Incremental efficiency estimates. Following section 8, in this subsection we assume that the optimal set S_* is nonempty, and that the sequences $\{x^k\}$, $\{\bar C_k\}$ (cf. (9.3)), and $\{\epsilon_k\}$ are bounded. Thus, replacing (8.4) by

$$\hat D := \sup_k d_{S_*}(x^k) \quad \text{and} \quad \hat G := \sup_k \bar C_k, \tag{9.22}$$

we have

(9.23)
$$\bar C_k := \sum_{i=1}^m \bar C_{ik} \le \hat G \le m\hat G_{\max} \quad \text{with} \quad |g_i^k| \le \bar C_{ik} \le \hat G_{\max} := \max_i \sup_k \bar C_{ik}.$$

We now give estimates for the Cesáro averages of the objective values $\bar f_k$ (cf. (8.1)), the Cesáro averages of the incremental objective values (cf. (9.1)) defined by

$$\bar f_{\rm inc}^k := \sum_{j=k'}^k \nu_j f_{\rm inc}^j / \nu_{\rm sum}^k \quad \text{with} \quad \nu_{\rm sum}^k := \sum_{j=k'}^k \nu_j, \tag{9.24}$$

and the objective values of the *incremental record points* (cf. [BTMN01, sect. 5])

$$\breve x^k := x^{\breve k} \quad \text{with} \quad \breve k \in \operatorname{Arg\,min}\{\, f_{\rm inc}^j : k' \le j \le k \,\}. \tag{9.25}$$

LEMMA 9.12. *In the notation of* (8.1), (9.23), (9.24), *and* (9.25), *we have*

$$\bar{f}_k - f_* \le \Delta_k + \bar{\epsilon}_k, \quad \Delta_k := \frac{d^2_{S_*}(x^{k'}) + \hat{G}^2 \sum_{j=k'}^k \nu_j^2}{2\sum_{j=k'}^k \nu_j}, \tag{9.26}$$

$$\bar{f}^k_{\mathrm{inc}} - f_* \le \bar{\Delta}_k + \bar{\epsilon}_k, \quad \bar{\Delta}_k := \frac{d^2_{S_*}(x^{k'}) + \min\{\hat{G}^2, m\hat{G}^2_{\max}\} \sum_{j=k'}^k \nu_j^2}{2\sum_{j=k'}^k \nu_j}, \tag{9.27}$$

$$f(\breve{x}^k) - f_* \le \breve{\Delta}_k + \bar{\epsilon}_k, \quad \breve{\Delta}_k := \bar{\Delta}_k + \frac{m-1}{2m}\hat{G}^2 \max_{j=k'\,:\,k} \nu_j. \tag{9.28}$$

Proof. Replace $|g^j|$ by $\bar{C}_j$ in (8.2) (cf. (9.5) and the proof of Corollary 9.2) and use (9.23) to get (9.26). Summing up (9.7) and using (9.24) and (8.1) (for $\bar{\epsilon}_k$) yields

$$\bar{f}^k_{\mathrm{inc}} - f_S(x) \le \frac{|x^{k'} - x|^2 + \sum_{j=k'}^k \nu_j^2 \sum_{i=1}^m |g_i^j|^2}{2\sum_{j=k'}^k \nu_j} + \bar{\epsilon}_k \quad \forall x. \tag{9.29}$$

Letting $x := P_{S_*}x^{k'}$ in (9.29) and bounding $\sum_i |g_i^j|^2 \le \min\{m\hat{G}^2_{\max}, \hat{G}^2\}$ (cf. (9.23)), we get (9.27). Next, we have $f^{\breve{k}}_{\mathrm{inc}} = \min_{j=k'}^k f^j_{\mathrm{inc}} \le \bar{f}^k_{\mathrm{inc}}$ by (9.24) and (9.25), whereas by (9.8) and (9.23)

$$f(\breve{x}^k) = f(x^{\breve{k}}) \le f^{\breve{k}}_{\mathrm{inc}} + \nu_{\breve{k}} \sum_{i=1}^m \bar{C}_{i\breve{k}} \sum_{j=1}^{i-1} \bar{C}_{j\breve{k}} \le f^{\breve{k}}_{\mathrm{inc}} + \nu_{\breve{k}} \tfrac{1}{2}\hat{G}^2(1 - \tfrac{1}{m})$$

(since $\sum_i \bar{C}^2_{i,\breve{k}} \ge \frac{1}{m}\bar{C}^2_{\breve{k}}$); combining these bounds with (9.27) gives (9.28). ∎

The estimate (9.26) bounds the objective values $f(\bar{x}^k) \le \bar{f}_k$ and $f(x^k_{\mathrm{rec}}) \le \bar{f}_k$ of the Cesáro points $\bar{x}^k$ and the record points x^k_{rec} (cf. (3.14), (8.3)).

We may now present efficiency estimates for stepsizes analogous to those of (8.9).

THEOREM 9.13. *Consider the following two stepsize rules and their efficiency factors:*

$$\nu_k := \frac{D_k k^{-s}}{G_k} \quad \textit{with} \quad c_{(9.30)} := G_{\max}\frac{\hat{D}^2 + D^2_{\max}(\hat{G}/G_{\min})^2}{D_{\min}}, \tag{9.30}$$

$$\nu_k := \frac{D_k k^{-s}}{mG_k} \quad \textit{with} \quad c_{(9.31)} := mG_{\max}\frac{\hat{D}^2 + D^2_{\max}(\hat{G}_{\max}/G_{\min})^2}{D_{\min}}, \tag{9.31}$$

where $s \in [1/2, 1]$, $\hat{D}$, $\hat{G}$ *and* $\hat{G}_{\max}$ *are defined by* (9.22)–(9.23), *and* D_k *and* G_k *are scaling factors that satisfy* (8.5). *Then for each rule we have for all* k

(9.32)

$$\bar{f}_k - f_* \le \bar{\epsilon}_k + \begin{cases} \dfrac{(1+\ln 2)c}{(4 - 2^{3/2})(k+1)^{1-s}} & \textit{if} \quad k' = \left\lceil \dfrac{1}{2}k \right\rceil, \\ \dfrac{\min\left\{\frac{2s}{2s-1}, 1 + \ln k\right\} c}{\max\left\{2\ln(k+1), (4 - 2^{3/2})(k+1)^{1-s}\right\}} & \textit{if} \quad k' = 1, \end{cases}$$

where $c := c_{(9.30)}$ *for the rule* (9.30) *and* $c := c_{(9.31)}$ *for the rule* (9.31). *Moreover, for the incremental record points* $\breve{x}^k$ *defined by* (9.25) *with* $k' = \lceil \frac{1}{2}k \rceil$, *we have for* *each* k

(9.33)
$$f(\breve{x}^k) - f_* \le \bar{\epsilon}_k + \frac{(1+\ln 2)c}{(4-2^{3/2})(k+1)^{1-s}} + \frac{D_{\max}}{2^{1-s}G_{\min}k^s}\begin{cases} \frac{m-1}{m}\hat{G}^2 & \textit{for} \quad (9.30), \\ (m-1)\hat{G}^2_{\max} & \textit{for} \quad (9.31), \end{cases}$$

$$f(\breve{x}^k) - f_* \le \bar{\epsilon}_k + \frac{(1+\ln 2)c}{(4-2^{3/2})k^{1/2}} \quad \textit{for} \quad s = 1/2, \tag{9.34}$$

where $c := \frac{3}{2}c_{(9.30)}$ *for the rule* (9.30) *and* $c := c_{(9.31)}$ *for the rule* (9.31)*. Further, if* $C_\epsilon := \sup_k k^s\epsilon_k$ *is finite, then the estimate* (8.11) *holds with* $c_\epsilon := \frac{D_{\max}G_{\max}}{D_{\min}G_{\min}}$ *so that* $\bar{\epsilon}_k$ *in* (9.32)*–*(9.34) *has the same order in* k *as its right neighbors.*

Proof. It suffices to bound Δ_k in (9.26) and $\breve{\Delta}_k$ in (9.28) by using $d_{S_*}(x^{k'}) \le \hat{D}$ (cf. (9.22)) and (8.5) together with Lemma 8.1 for the sums. ∎

Remark 9.14.

(i) For both stepsize rules (9.30)–(9.31), D_k should be a guess for $d_{S_*}(x^k)$ (or for the "diameter of the picture"), but for the first one G_k should be a guess for $\hat{G}$ (e.g., $\sum_i |g_i^{k-1}|$), whereas for the second one G_k should be a guess for $\hat{G}_{\max}$ (e.g., $\max_i |g_i^{k-1}|$).

(ii) For comparisons, suppose the feasible set S is bounded and the subgradients of each objective f_i are exact ($\epsilon_i^k \equiv 0$) and bounded by its Lipschitz constant L_{f_i} on S so that $\hat{D}$ may be replaced by $\operatorname{diam}(S)$, $\hat{G}$ by $\sum_i L_{f_i}$, and $\hat{G}_{\max}$ by $\max_i L_{f_i}$. Further, assume that $D_{\min}$ and $D_{\max}$ are of order $\hat{D}$, $G_{\min}$ and $G_{\max}$ are of order $\hat{G}$ for (9.30) and $\hat{G}_{\max}$ for (9.31) so that $c_{(9.30)} \approx 2\operatorname{diam}(S)\sum_i L_{f_i}$ and $c_{(9.31)} \approx 2\operatorname{diam}(S) m \max_i L_{f_i}$. Under similar assumptions, the nonincremental version has $c_{(8.9)} \approx 2\operatorname{diam}(S)L_f$, where L_f is the Lipschitz constant of f on S. Of course, $L_f \le \sum_i L_{f_i} \le m \max_i L_{f_i}$. Assuming that $\max_i L_{f_i} \le L_f$ (as in [BTMN01, Thm. 5.1]), the efficiency estimates for the incremental version given in Theorem 9.13 are at most m times larger than those for the ordinary version stated in Theorem 8.2; yet their ratio decreases when $\sum_i L_{f_i}$ gets closer to L_f; i.e., all f_i become "similar." Such "similarity" features help the incremental version to be competitive in practice [BTMN01, NeB01].

(iii) Remark 8.3(i) on the choice of s and k' remains valid.

(iv) In the exact case of $\epsilon_k \equiv 0$, our estimate (9.34) for the stepsize rule (9.31) is similar to that of [BTMN01, Thm. 5.1] (for the Euclidean norm).

For nonvanishing stepsizes $\nu_k \equiv \nu$, the asymptotic objective accuracy is of order $\frac{1}{2}\hat{G}^2\nu \le \frac{1}{2}m^2\hat{G}^2_{\max}\nu$ (cf. Corollary 9.2, Thm. 3.2, and (9.22)–(9.23)), and the relative accuracy may be estimated as in Proposition 8.4 (cf. (8.17)).

Proposition 9.15. *For a fixed stepsize* $\nu_k \equiv \nu > 0$*, we have the following efficiency bounds on* Δ_k *and* $\breve{\Delta}_k$ *defined by* (9.26) *and* (9.28) *with* $k' = 1$:

$$\Delta_k \le \frac{1}{2}\hat{G}^2\nu\left(1 + \frac{\hat{G}^2D^2}{(\hat{G}^2\nu)^2k}\right), \tag{9.35}$$

$$\breve{\Delta}_k \le \frac{1}{2}\hat{G}^2\nu\left(1 + \frac{m-1}{m} + \frac{\hat{G}^2D^2}{(\hat{G}^2\nu)^2k}\right), \tag{9.36}$$

$$\max\left\{\Delta_k, \breve{\Delta}_k\right\} \le \frac{1}{2} m^2 \hat{G}_{\max}^2 \nu \left(1 + \frac{m^2 \hat{G}_{\max}^2 D^2}{(m^2 \hat{G}_{\max}^2 \nu)^2 k}\right), \tag{9.37}$$

where $D := d_{S_*}(x^1)$, $\hat{G}_{\max} := \sup_{i,k} \bar{C}_{ik}$, *and* $\hat{G} := \sup_k \bar{C}_k \le m\hat{G}_{\max}$.

Proof. This follows easily from the definitions (9.26) and (9.28). ∎

Acknowledgments. I would like to thank the Associate Editor and the two anonymous referees for helpful comments. Further, I would like to acknowledge extensive discussions with Ya. I. Alber, D. Bertsekas, O. Burdakov, C. Lemaréchal, and A. Nedić.

REFERENCES

[AIS98] Ya. I. Alber, A. N. Iusem, and M. V. Solodov, *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space*, Math. Programming, 81 (1998), pp. 23–35.

[Ber97] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.

[Ber99] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[BeT00] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.

[Brä93] U. Brännlund, *On Relaxation Methods for Nonsmooth Convex Optimization*, Ph.D. thesis, Department of Mathematics, Royal Institute of Technology, Stockholm, 1993.

[Brä95] U. Brännlund, *A generalized subgradient method with relaxation step*, Math. Programming, 71 (1995), pp. 207–219.

[BSS93] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., Wiley, New York, 1993.

[BTMN01] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108.

[CoL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[DeV81] V. F. Demyanov and L. V. Vasilev, *Nondifferentiable Optimization*, Nauka, Moscow, 1981 (in Russian); Optimization Software Inc., New York, 1985 (in English).

[DuS88] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I*: General Theory*, Wiley-Interscience, New York, 1988.

[Erm66] Yu. M. Ermoliev, *Methods of solution of nonlinear extremal problems*, Kibernetika, no. 4 (1966), pp. 1–17 (in Russian); Cybernetics, 2 (1966), pp. 1–16 (in English).

[Erm76] Yu. M. Ermoliev, *Stochastic Programming Methods*, Nauka, Moscow, 1976 (in Russian).

[ErS68] Yu. M. Ermoliev and N. Z. Shor, *A random search method for a two-stage stochastic programming problem and its generalization*, Kibernetika, no. 1 (1968), pp. 90–92 (in Russian).

[Gai94] A. A. Gaivoronski, *Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part* 1, Optim. Methods Softw., 4 (1994), pp. 117–134.

[Gla65] E. G. Gladishev, *On stochastic approximation*, Theory Probab. Appl., 10 (1965), pp. 297–300 (in Russian).

[GoK99] J.-L. Goffin and K. C. Kiwiel, *Convergence of a simple subgradient level method*, Math. Program., 85 (1999), pp. 207–211.

[Gri94] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Softw., 4 (1994), pp. 135–150.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[KiA91] S. Kim and H. Ahn, *Convergence of a generalized subgradient method for nondifferentiable optimization*, Math. Programming, 50 (1991), pp. 75–80.

[Kib79] V. M. Kibardin, *Decomposition into functions in the minimization problem*, Avtomat. i Telemekh., no. 9 (1979), pp. 66–79 (in Russian); Automat. Remote Control, 40 (1980), pp. 1311–1323 (in English).

[KiU93] S. Kim and B.-S. Um, *An improved subgradient method for constrained nondifferentiable optimization*, Oper. Res. Lett., 14 (1993), pp. 61–64.

[Kiw96a] K. C. Kiwiel, *The efficiency of subgradient projection methods for convex optimization, Part* II*: Implementations and extensions*, SIAM J. Control Optim., 34 (1996), pp. 677–697.

[Kiw96b] K. C. Kiwiel, *A Subgradient Method with Bregman Projections for Convex Constrained Nondifferentiable Minimization*, Tech. report, Systems Research Institute, Warsaw, Poland, 1996.

[Kiw98] K. C. Kiwiel, *Subgradient method with entropic projections for convex nondifferentiable minimization*, J. Optim. Theory Appl., 96 (1998), pp. 159–173.

[KLL99a] K. C. Kiwiel, T. Larsson, and P. O. Lindberg, *Dual Properties of Ballstep Subgradient Methods, with Applications to Lagrangian Relaxation*, Tech. report LiTH-MAT-R-1999-24, Department of Mathematics, Linköping University, Linköping, Sweden, 1999. Revised 2002.

[KLL99b] K. C. Kiwiel, T. Larsson, and P. O. Lindberg, *The efficiency of ballstep subgradient level methods for convex optimization*, Math. Oper. Res., 24 (1999), pp. 237–254.

[Lis86] S. A. Lisina, *A subgradient method for minimizing a convex function for the case when the infimum is not achieved*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom., no. 4 (1986), pp. 70–74 (in Russian).

[Lit68] B. M. Litvakov, *Convergence of recurrent algorithms for pattern recognition learning*, Avtomat. i Telemekh., no. 3 (1968), pp. 142–150 (in Russian); Automat. Remote Control, 29 (1968), pp. 121–128 (in English).

[LPS96] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization - theory and applications*, European J. Oper. Res., 88 (1996), pp. 382–403.

[LPS00] T. Larsson, M. Patriksson, and A.-B. Strömberg, *On the Convergence of Conditional ϵ-Subgradient Methods for Convex Programs and Convex-Concave Saddle-Point Problems*, Tech. report, Department of Mathematics, Chalmers University of Technology, Linköping, Sweden, 2000.

[Luo91] Z. Q. Luo, *On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks*, Neural Computation, 3 (1991), pp. 226–245.

[LuT94] Z.-Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw., 4 (1994), pp. 85–101.

[MaS94] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Softw., 4 (1994), pp. 103–116.

[MGN87] V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin, *Methods of Nonconvex Optimization*, Nauka, Moscow, 1987 (in Russian).

[Min86] M. Minoux, *Mathematical Programming, Theory and Algorithms*, John Wiley and Sons, Chichester, UK, 1986.

[MiU82] F. Mirzoakhmedov and S. P. Uryasev, *Adaptive stepsize control for the stochastic approximation algorithm*, Zh. Vychisl. Mat. i Mat. Fiz., 23 (1982), pp. 1314–1325 (in Russian).

[NeB01] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.

[Nes84] Yu. E. Nesterov, *Minimization methods for nonsmooth convex and quasiconvex functions*, Èkonom. i Mat. Metody, 20 (1984), pp. 519–531 (in Russian); Matekon, 29 (1984), pp. 519–531 (in English).

[Nes89] Yu. E. Nesterov, *Effective Methods in Nonlinear Programming*, Radio i Sviaz, Moscow, 1989 (in Russian).

[NeY78] A. S. Nemirovskii and D. B. Yudin, *Cesaro convergence of the gradient method for approximating saddle points of convex-concave functions*, Dokl. Akad. Nauk SSSR, 239 (1978), pp. 1056–1059 (in Russian).

[Nur79] E. A. Nurminskii, *Numerical Methods for Solving Deterministic and Stochastic Minimax Problems*, Naukova Dumka, Kiev, 1979 (in Russian).

[Nur82] E. A. Nurminski, *Subgradient method for minimizing weakly convex functions and ϵ-subgradient methods of convex optimization*, in Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., CP–82–S8, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982, pp. 97–123.

[Nur91] E. A. Nurminskii, *Numerical Methods for Convex Optimization*, Nauka, Moscow, 1991 (in Russian).

[NuZ77] E. A. Nurminskii and A. A. Zhelikhovskii, *ϵ-Quasigradient method for solving non-*

smooth extremal problems, Kibernetika, no. 1 (1977), pp. 109–113 (in Russian); Cybernetics, 13 (1977), pp. 109–114 (in English).

[Pol67] B. T. Polyak, *A general method for solving extremum problems*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 33–36 (in Russian); Soviet Math. Dokl., 8 (1967), pp. 593–597 (in English).

[Pol69] B. T. Polyak, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 509–521 (in Russian); U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 14–29 (in English).

[Pol78] B. T. Polyak, *Subgradient methods: A survey of Soviet research*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, UK, 1978, pp. 5–29.

[Pol83] B. T. Polyak, *Introduction to Optimization*, Nauka, Moscow, 1983 (in Russian); Optimization Software Inc., New York, 1987 (in English).

[PoT73] B. T. Polyak and Ya. Z. Tsypkin, *Pseudogradient adaptation and training algorithms*, Avtomat. i Telemekh., no. 3 (1973), pp. 45–68 (in Russian); Automat. Remote Control, 34 (1973), pp. 377–397 (in English).

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Sch83] K. Schulz, *A note on the convergence of subgradient optimization methods*, Math. Operationsforsch. Statist. Ser. Optim., 14 (1983), pp. 537–541.

[SCT00] H. D. Sherali, G. Choi, and C. H. Tuncbilek, *A variable target value method for nondifferentiable optimization*, Oper. Res. Lett., 26 (2000), pp. 1–8.

[Sho62] N. Z. Shor, *An application of the generalized gradient descent method to the solution of a network transportation problem*, in Proceedings of the Scientific Seminar on Theoretical and Application Problems of Cybernetics and Operations Research, no. 1, Scientific Council on Cybernetics of the Academy of Sciences of the Ukrainian SSR, Kiev, 1962, pp. 9–17 (in Russian).

[Sho79] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Naukova Dumka, Kiev, 1979 (in Russian); Springer-Verlag, Berlin, 1985 (in English).

[ShW96] A. Shapiro and Y. Wardi, *Convergence analysis of gradient descent stochastic algorithms*, J. Optim. Theory Appl., 91 (1996), pp. 439–454.