SIAM J. OPTIM.
Vol. 20, No. 1, pp. 357–386

CONVERGENCE ANALYSIS OF DEFLECTED CONDITIONAL APPROXIMATE SUBGRADIENT METHODS*

GIACOMO D'ANTONIO† AND ANTONIO FRANGIONI‡

Abstract. Subgradient methods for nondifferentiable optimization benefit from *deflection*, i.e., defining the search direction as a combination of the previous direction and the current subgradient. In the constrained case they also benefit from *projection* of the search direction onto the feasible set prior to computing the steplength, that is, from the use of conditional subgradient techniques. However, combining the two techniques is not straightforward, especially if an *inexact oracle* is available which can only compute approximate function values and subgradients. We present a convergence analysis of several different variants, both conceptual and implementable, of approximate conditional deflected subgradient methods. Our analysis extends the available results in the literature by using the main stepsize rules presented so far, while allowing deflection in a more flexible way. Furthermore, to allow for (diminishing/square summable) rules where the stepsize is tightly controlled a priori, we propose a new class of deflection-restricted approaches where it is the deflection parameter, rather than the stepsize, which is dynamically adjusted using the "target value" of the optimization sequence. For both Polyak-type and diminishing/square summable stepsizes, we propose a "correction" of the standard formula which shows that, in the inexact case, knowledge about the error computed by the oracle (which is available in several practical applications) can be exploited in order to strengthen the convergence properties of the method. The analysis allows for several variants of the algorithm; at least one of them is likely to show numerical performances similar to these of "heavy ball" subgradient methods, popular within backpropagation approaches to train neural networks, while possessing stronger convergence properties.

Key words. convex programming, nondifferentiable optimization, subgradient methods, convergence analysis, Lagrangian relaxation, backpropagation

AMS subject classification. 90C25

DOI. 10.1137/080718814

1. Introduction.

We are concerned with the numerical solution of the nondifferentiable optimization (NDO) problem

$$f^* = \inf \{ f(x) : x \in X \}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is finite-valued and convex (hence, continuous) and $X \subseteq \mathbb{R}^n$ is closed convex. We are specifically interested in the case where $X \neq \mathbb{R}^n$, that is, (1.1) is a *constrained* NDO problem; X has to be given in a form that allows easy projection (in most applications X is a very simple polyhedron, such as the nonnegative orthant). It is customary to assume that f is only known through an oracle ("black box") that, given any $x \in X$, returns the value $f(x)$ and one *subgradient* $g \in \partial f(x)$. In order to make the algorithm more readily implementable, it is useful to contemplate the case where only an "approximate" subgradient can be obtained; that is, we will allow g to only satisfy the "relaxed" subgradient inequality $f(y) \geq f(x) + \langle g, y - x \rangle - \varepsilon$ for all y and some $\varepsilon \geq 0$, i.e., to belong to the (larger) ε-subdifferential of f at x, denoted by $\partial_\varepsilon f(x)$. This is particularly useful in *Lagrangian relaxation*, where the "black box"

*Received by the editors March 18, 2008; accepted for publication (in revised form) December 17, 2008; published electronically April 24, 2009.
http://www.siam.org/journals/siopt/20-1/71881.html
†Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy (dantonio@dm.unipi.it).
‡Dipartimento di Informatica, Università di Pisa, Polo Universitario della Spezia, Via dei Colli 90, 19121 La Spezia, Italy (frangio@di.unipi.it).

requires the approximate solution of a potentially difficult optimization subproblem. Since in this case ε is precisely the absolute tolerance required, allowing for a "large" ε may substantially decrease the oracle time required.

We will study solution algorithms for (1.1) belonging to the class of *subgradient methods*. These algorithms, introduced by Polyak in his seminal paper from 1969 (see [31] for a review of the early contributions in the field) have been for a long time the only computationally viable approach for solving (1.1). Despite the emergence of other classes of algorithms such as bundle [18, 14] and centers-based methods [11, 28], which are often more efficient, subgradient approaches may still be a valuable alternative, especially for very-large-scale problems and if the required accuracy for the solution is not too high [9, 12].

We consider subgradient methods for (1.1) based on the recurrence equation

$$\widehat{x}_{k+1} = x_k - \nu_k d_k, \quad x_{k+1} = P_X(\widehat{x}_{k+1}), \tag{1.2}$$

where P_X denotes orthogonal projection on X and d_k is the (opposite of the) *search direction*, computed using the *current (approximate) subgradient* $g_k \in \partial_{\varepsilon_k} f(x_k)$ and possibly information from the previous iteration, while $\nu_k \geq 0$ is the *stepsize*. While the original subgradient methods [30] used $d_k = g_k$, it soon became clear that some form of *deflection*, i.e., using $d_k = g_k + \eta_k v_k$, was very important in order to improve practical performances. One idea was to use $v_k = d_{k-1}$, with η_k chosen in such a way that $\langle d_k, d_{k-1} \rangle \geq 0$ [7]; this "dampens" the zig-zagging phenomenon whereby the direction at one step is almost opposite to that of the previous step, yielding very slow convergence. Another approach is that of "heavy ball" subgradient methods [30, 35], which rather use $v_k = x_k - x_{k-1}$ (called *momentum term*); while d_{k-1} and $x_k - x_{k-1}$ are collinear in the unconstrained case ($X = \mathbb{R}^n$), this is *not* so in general due to the projection.

Deflection does not, however, cure the other form of zig-zagging, which occurs when the iterates x_k are on the boundary of X and the directions d_k turn out to be almost orthogonal to the frontier. This is a consequence of the fact that, in the original subgradient iteration, the feasible set X is *not* considered during the construction of the direction d_k, but only a posteriori. A possible remedy is readily obtained by simply considering the *essential objective* $f_X(x) = f(x) + I_X(x)$ of (1.1), I_X being the *indicator function* of X. A (ε-)subgradient of f_X is said to be a *conditional (ε)-subgradient* of f w.r.t. X [22, 24] and can be used instead of g_k to compute the search direction. It is well known that $\partial_\varepsilon f_X(x) \supseteq \partial_\varepsilon f(x) + \partial I_X(x) = \partial_\varepsilon f(x) + N_k$ [18], where $N_k = N_X(x_k)$ is the *normal cone* of X at x_k; thus, for iterates x_k on the frontier of X (where $N_k \neq \{0\}$), one may have, for a given g_k produced by the "black box," multiple choices of vectors in the normal cone to produce a conditional subgradient $\widehat{g}_k$ to be used for computing the direction. The obvious choice is to select the optimal solution of

$$\operatorname{argmin} \left\{ \| g \|^2 \ : \ g \in g_k + N_k \right\},$$

which, if f happens to be differentiable at x_k and $\varepsilon_k = 0$, is the steepest descent direction; hence, in the (unlikely) case that $\widehat{g}_k = 0$, one would have proven optimality of x_k. Denoting by $T_k = T_X(x_k)$ the *tangent cone* of X at x_k, it is well known that T_k and N_k are *polar cones*, that is, $\langle v, w \rangle \leq 0$ for each $v \in T_k$ and $w \in N_k$ (which is, in particular, true when x_k is in the interior of X, so that $T_k = \mathbb{R}^n$ and $N_k = \{0\}$); thus, by the *Moreau decomposition* principle, $\widehat{g}_k$ is also the solution of

$$\widehat{g}_k = \operatorname{argmin} \left\{ \| g - g_k \|^2 \ : \ g \in -T_k \right\} = -P_{T_k}(-g_k). \tag{1.3}$$

This gives rise to the *projected subgradient* approach, where $d_k = -P_{T_k}(-g_k)$. Convergence of approaches using conditional (ε-)subgradients can be proven under common assumptions on the stepsize [22, 24, 20].

However, to the best of our knowledge, no explicit convergence proof is known for subgradient methods which combine these two techniques. In [17], a "hybrid" subgradient approach is proposed which employs deflection when x_k lies in the interior of X and projection when x_k is on the boundary; however, one would clearly prefer to be able to deflect at every iteration. The issue here is that projecting the subgradient and deflecting simultaneously, i.e., using $d_k = \widehat{g}_k + \eta_k v_k$, with the above choices of v_k, would hardly result in an efficient approach since d_k is unlikely to belong to T_k; thus, even in the polyhedral case the approach would not produce feasible directions. Our development is based on two main ideas: first, we restrict ourselves to deflection formulae akin to

$$d_k = \alpha_k g_k + (1-\alpha_k) v_k, \quad \alpha_k \in [0,1], \tag{1.4}$$

i.e., where the direction is taken as a convex combination of the previous direction and the current subgradient. This choice can be motivated as follows:

- Any deflection rule where g_k is not scaled can be seen as (1.4), where d_k is afterwards scaled by $1/\alpha_k$, an effect that can alternatively be taken into account by changing the stepsize ν_k; for instance, a simple condition over α_k in (1.4) might be used to ensure that $\langle d_k, d_{k-1}\rangle \geq 0$, the original aim of [7].
- Choice (1.4) guarantees that $v_k \in \partial_{\varepsilon'_k} f_X(x_k)$ for a proper ε'_k—*different* from the oracle error ε_k—which can be explicitly computed (Lemma 2.6) and controlled (Lemmas 3.1 and 4.1), thereby allowing to exploit known results [8, 20].
- This is the choice of *volume-like* variants of the approach [3, 2, 34], which have become popular in the important Lagrangian application [25, 16, 15] due to their ability to (asymptotically) provide *primal optimal solutions* to the "convexified relaxation" (although this is not the only means to extract primal solutions out of a subgradient algorithm [23, 1, 29, 32]).

The second key idea is that, since (1.4) guarantees that the direction is a ε'_k-subgradient, instead of projecting g_k, one may (and should) choose to *project* d_k. A closer inspection reveals that there are actually *eight* different ways in which this can be done:

$$\bar{g}_k \in \{\, g_k \,,\, \widehat{g}_k \,\}, \quad v_k \in \left\{\, \tilde{d}_{k-1} \,,\, \widehat{d}_{k-1} \,\right\}, \tag{1.5}$$

$$\tilde{d}_k = \alpha_k \bar{g}_k + (1-\alpha_k) v_k, \quad d_k \in \left\{\, \tilde{d}_k \,,\, \widehat{d}_k = -P_{T_k}\left(-\tilde{d}_k\right) \,\right\}, \tag{1.6}$$

where we assume $\alpha_1 = 1$, thus rendering v_1 irrelevant and the formulae well defined. That is, at each step, one has two (approximate) subgradients g_k and v_k, each of which can be individually deflected (or not); their convex combination $\tilde{d}_k$ is formed, and either it is directly used as the direction, or it is projected beforehand. Of course, the combination where no projection is ever performed, apart from that of $\widehat{x}_k$, can hardly be considered a conditional subgradient approach. Similarly, the other three where $d_k = \tilde{d}_k$ is *not* projected may be suspected to suffer the zig-zagging phenomenon, as even projection of g_k and use of $v_k = \widehat{d}_{k-1}$—which is, indeed, projected, but w.r.t. the previous iterate x_{k-1}—cannot guarantee that $\tilde{d}_k \in T_k$. However, even restricting to the four cases where $d_k = \widehat{d}_k$, the simple example $g_k = (1,-1)$, $v_k = (-1,1)$, $x_k = (0,0)$, $X = \mathbb{R}^2_+$, and $\alpha_k = 1/2$ shows that the four schemes can provide four different directions. Actually, it appears that *the treatment could be extended to allowing* v_k *to*

be any convex combination of $\tilde{d}_{k-1}$ *and* $\widehat{d}_{k-1}$ and similarly for $\bar{g}_k$, but there seems to be little reason (or sensible way) to choose anything but the extreme cases. Also, note that the selection between the different projection alternatives can vary arbitrarily at each iteration.

In this article we present unified convergence results for *conditional, deflected, approximate* subgradient algorithms of the form (1.5)/(1.6). We consider both a modified version of Polyak stepsize, and diminishing/square summable ones; starting from "abstract" rules requiring the knowledge of f^*, we work our way towards "concrete" ones which do away with this condition, as originally proposed by Polyak himself [30]. The convergence analysis is centered on the fact that the *deflection term* v_k is, at every iteration, a conditional ε'_k-subgradient for a ε'_k that is, in general, different from the "oracle error" of the current iterate; hence, the proposed approach is a conditional approximate subgradient method, and as such it falls under the very general study of [20]. However, in this particular case, the accuracy of the subgradient (ε'_k) depends in a complex way on the deflection parameter (α_k) and the specific projection formula used. We, therefore, provide implementable rules for the selection of α_k which ensure convergence of the approach, as opposed to providing abstract conditions which are somewhat required to hold. Furthermore, our analysis of *corrected* rules improves on the available convergence results even for nondeflected approaches.

2. Preliminary results. In the following, we will always assume a (conditional, deflected, approximate) subgradient method of the form (1.2) with direction chosen by (1.5)/(1.6). We will denote by $X^* \subseteq X$ the optimal solution set of (1.1) and by x^* any one of its elements; that is, $f^* = f(x^*) > -\infty$ if $X^* \neq \emptyset$. We will assume that

$$g_k \in \partial_{\sigma_k} f(x_k), \tag{2.1}$$

i.e., we will denote by $\sigma_k \geq 0$ the "oracle error," leaving the notation ε_k for the "direction error" (cf. Lemma 2.6), in order to notationally simplify the comparison with known results [8, 20]. Finally, we denote by $f_k = f(x_k)$ the sequence of function values and by $f^\infty = \liminf_{k\to\infty} f_k$ its asymptotic value.

For the standard subgradient method (i.e., $d_k = g_k$ and $\sigma_k = 0$), one way of ensuring convergence is to guarantee the "nonexpansivity" property

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|, \tag{2.2}$$

which is implied by

$$\langle -d_k, x^* - x_k \rangle \geq 0, \tag{2.3}$$

which, in turn, immediately follows from the definition of $d_k = g_k \in \partial f(x_k)$, i.e.,

$$\langle -d_k, x_k - x^* \rangle = \langle d_k, x^* - x_k \rangle \leq f(x^*) - f_k \leq 0. \tag{2.4}$$

However, when $d_k \neq g_k$, these properties fail to hold unless the stepsize ν_k and the deflection coefficient α_k are properly managed. This is illustrated in Figure 2.1.

The crucial property (2.3) requires "moving in a right direction," i.e., that $X^* \subseteq \{ x \in \mathbb{R}^n : \langle -d_k, x - x_k \rangle \geq 0 \}$. For $v_k = d_{k-1}$, (2.3) is implied by

$$\langle d_{k-1}, x^* - x_k \rangle \leq 0. \tag{2.5}$$

Therefore, one may impose (2.3) *irrespective of* α_k by requiring the stepsize to be small enough; in the figure, $x^{(a)}_{k+1}$ is the point farthest from x_k which satisfies this

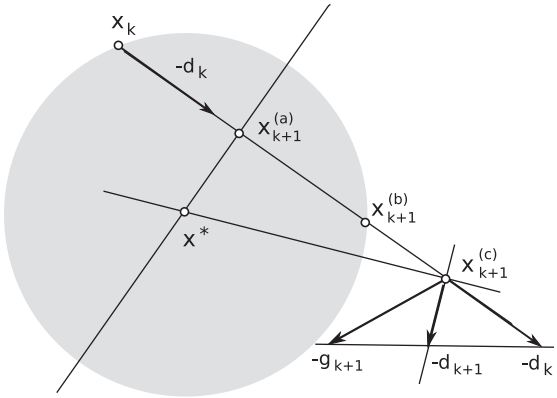

FIG. 2.1. *Relationships between* ν_k, α_k, *and* (2.3).

condition. However, (2.2) is satisfied by a possibly larger set of stepsizes; in the figure, this is represented by point $x_{k+1}^{(b)}$. If, instead, one is wary about limiting the stepsize and, say, obtains point $x_{k+1}^{(c)}$ in the figure, it is still possible to ensure that (2.3) holds by imposing that d_k is "not too different" from g_k, i.e., imposing some lower bound on α_k; this corresponds to direction d_{k+1} in the figure.

Summarizing, in order to ensure that sufficient conditions for convergence hold, either the stepsize has to be properly limited in order to allow any deflection, or the deflection has to be properly limited in order to allow any stepsize. Therefore, in the following we will separately study two different kinds of subgradient schemes: *stepsize-restricted* approaches and *deflection-restricted* ones. Before doing that, we provide some technical lemmas that are useful for both.

2.1. Technical lemmas. We start recalling a few known results about the geometry of the involved points and directions.

LEMMA 2.1. *Let* $p_k = x_k - \widehat{x}_k$; *then*

$$\langle -p_{k+1}, d_k \rangle \le 0, \tag{2.6}$$

$$\|x_{k+1} - x_k\| \le \|\widehat{x}_{k+1} - x_k\|, \tag{2.7}$$

$$\langle p_{k+1}, x_{k+1} - x_k \rangle \le 0. \tag{2.8}$$

Proof. For (2.6) and (2.7), see [17, Lemma 3.9]; for (2.8), see [5, Proposition 2.2.1]. ∎

LEMMA 2.2. *For any* $x \in X$, *one has*

$$\|x_{k+1} - x\|^2 \le \|x_k - x\|^2 - 2\nu_k \langle x_k - x, d_k \rangle + \nu_k^2 \|d_k\|^2. \tag{2.9}$$

Proof. Using [5, Theorem 2.2.1] for the first step, one has

$$\begin{aligned} \|x_{k+1} - x\|^2 \le \|\widehat{x}_{k+1} - x\|^2 &= \|x_k - \nu_k d_k - x\|^2 \\ &= \|x_k - x\|^2 - 2\nu_k \langle x_k - x, d_k \rangle + \nu_k^2 \|d_k\|^2. \end{aligned}$$ ∎

We now proceed with the main technical results, aimed at reproducing (2.2)–(2.4) in the more general deflected conditional setting.

LEMMA 2.3. *For each* $x \in X$, *it holds*

$$\left\langle \widehat{d}_k, x - x_k \right\rangle \le \langle d_k, x - x_k \rangle \le \left\langle \tilde{d}_k, x - x_k \right\rangle. \tag{2.10}$$

Proof. From $-\tilde{d}_k = P_{T_k}(-\tilde{d}_k) + P_{N_k}(-\tilde{d}_k) = -\widehat{d}_k + P_{N_k}(-\tilde{d}_k)$ one obtains

$$\left\langle \tilde{d}_k, x - x_k \right\rangle = \left\langle \widehat{d}_k, x - x_k \right\rangle - \left\langle P_{N_k}\left(-\tilde{d}_k\right), x - x_k \right\rangle \geq \left\langle \widehat{d}_k, x - x_k \right\rangle$$

from $P_{N_k}(-\tilde{d}_k) \in N_k$, $(x - x_k) \in T_k$. The rest follows from (1.6). ∎

An immediate consequence of (2.10) is that for any $x \in X$, one has

$$\langle v_{k+1}, x - x_k \rangle \leq \left\langle \tilde{d}_k, x - x_k \right\rangle. \tag{2.11}$$

In all but one case, (2.11) can be complemented by

$$\langle d_k, x - x_k \rangle \leq \langle v_{k+1}, x - x_k \rangle. \tag{2.12}$$

This property is crucial for convergence of the stepsize-restricted approaches, for reasons discussed below; the following lemma proves that it is always satisfied when $d_k = \widehat{d}_k$, i.e., in the "truly conditional" variants.

LEMMA 2.4. *Condition* (2.12) *holds if*

$$d_k = \tilde{d}_k \quad \Rightarrow \quad v_{k+1} = \tilde{d}_k. \tag{2.13}$$

Proof. Clearly, (2.12) holds if $d_k = v_{k+1}$; it also holds if $d_k = \widehat{d}_k$ and $v_{k+1} = \tilde{d}_k$ due to (2.10). Thus, the only "bad case" is when $d_k = \tilde{d}_k$ and $v_{k+1} = \widehat{d}_k$. ∎

LEMMA 2.5. *It holds*

$$\left\langle \tilde{d}_k, x_k - x_{k+1} \right\rangle \leq \langle d_k, x_k - x_{k+1} \rangle \leq \nu_k \|d_k\|^2. \tag{2.14}$$

Proof. For the rightmost inequality in (2.14), we have

$$\begin{aligned} \langle d_k, x_k - x_{k+1} \rangle &= \langle d_k, x_k - \widehat{x}_{k+1} \rangle + \langle d_k, \widehat{x}_{k+1} - x_{k+1} \rangle \qquad \text{(for (1.2))} \\ &= \nu_k \|d_k\|^2 + \langle d_k, \widehat{x}_{k+1} - x_{k+1} \rangle \leq \nu_k \|d_k\|^2 \quad \text{(for (2.6))}. \end{aligned}$$

The leftmost inequality comes from (2.10) with $x = x_{k+1}$. ∎

In the spirit of (2.11), (2.14) can be rewritten as

$$\langle v_{k+1}, x_k - x_{k+1} \rangle \leq \nu_k \|d_k\|^2. \tag{2.15}$$

LEMMA 2.6. *At all iterations* k, $v_{k+1} \in \partial_{\varepsilon_k} f_X(x_k)$, *with*

$$0 \leq \varepsilon_k = (1 - \alpha_k)\left(f_k - f_{k-1} - \langle v_k, x_k - x_{k-1} \rangle + \varepsilon_{k-1} \right) + \alpha_k \sigma_k. \tag{2.16}$$

Proof. The proof is by induction over k. For $k = 1$, $\alpha_1 = 1$ in (1.6)/(1.5) implies $v_2 = \bar{g}_1 \in \partial_{\sigma_1} f_X(x_1)$ ($\varepsilon_1 = \sigma_1 \geq 0$). For the inductive step $k > 1$, first observe that $v_k \in \partial_{\varepsilon_{k-1}} f_X(x_{k-1})$ immediately implies that

$$\varepsilon'_k = f_k - f_{k-1} - \langle v_k, x_k - x_{k-1} \rangle + \varepsilon_{k-1} \geq 0, \tag{2.17}$$

and therefore, $\varepsilon_k \geq 0$. Then, consider any fixed $x \in X$; from (1.6)

$$\begin{aligned} \left\langle \tilde{d}_k, x - x_k \right\rangle &= \alpha_k \langle \bar{g}_k, x - x_k \rangle + (1 - \alpha_k) \langle v_k, x - x_k \rangle \\ &= \alpha_k \langle \bar{g}_k, x - x_k \rangle + (1 - \alpha_k)\left(\langle v_k, x - x_{k-1} \rangle - \langle v_k, x_k - x_{k-1} \rangle \right). \end{aligned}$$

Applying (2.11), the inductive hypothesis, and $\bar{g}_k \in \partial_{\sigma_k} f_X(x_k)$, we then obtain

$$\begin{aligned}\langle v_{k+1}, x - x_k\rangle &\le \left\langle \tilde{d}_k, x - x_k \right\rangle \\ &\le \alpha_k(f(x) - f_k + \sigma_k) + (1-\alpha_k)\left(f(x) - f_{k-1} + \varepsilon_{k-1} - \langle v_k, x_k - x_{k-1}\rangle\right) \\ &= f(x) - f_k + (1-\alpha_k)\left(\, f_k - f_{k-1} - \langle v_k, x_k - x_{k-1}\rangle + \varepsilon_{k-1}\,\right) + \alpha_k\sigma_k\end{aligned}$$

(in the last line, we have added and subtracted $(1-\alpha_k)f_k$). □

A few remarks about Lemma 2.6 are in order.

1. The lemma proves the anticipated result that the deflection term v_k is an approximate subgradient of f_X. While in (2.16) this is true *for the previous iterate* x_{k-1}, the well-known *information transport property* ensures that any approximate subgradient at x_{k-1} is also an approximate subgradient at x_k, only with a different approximation. Indeed, from $v_k \in \partial_{\varepsilon_{k-1}} f_X(x_{k-1})$ it is easy to establish that $v_k \in \partial_{\varepsilon'_k} f_X(x_k)$, where ε'_k is given in (2.17). This (conditional) ε'_k-subgradient is combined with the new σ_k-subgradient (conditional or not) at x_k to obtain the final $\tilde{d}_k$, which then is a (conditional) ε_k-subgradient at x_k for $\varepsilon_k = (1-\alpha_k)\varepsilon'_k + \alpha_k\sigma_k$ (cf. (2.16)). Because N_k is a cone, further projecting $\tilde{d}_k$ keeps it in the conditional subdifferential.
2. *If σ_k is known*, then both ε_k and ε'_k are *computable*, i.e., they can easily be kept updated during the algorithm's operations. Thus, the natural (although seldom effective in practice) early stopping rule based on checking whether $||g_k||$ (or $||\bar{g}_k||$) and σ_k are "small" can be complemented with the analogous one checking $||v_k||$ and ε'_k. Even better yet, this may be used to *select* α_k. For instance, the value of α_k which minimizes $\|\tilde{d}_k\|$ can be found by a closed formula, while the one which minimizes $\|\widehat{d}_k\|$ requires a more costly constrained quadratic problem, albeit a specially structured one [13]. This would bring the subgradient algorithm very close to a Bundle method with "minimal" bundle [18, 14], a la [2].
3. The above discussion proves that *if* (2.13) *holds, then* $d_k \in \partial_{\varepsilon_k} f_X(x_k)$, i.e., the approach is a (conditional) ε_k-subgradient algorithm, where ε_k is given by the (complicated) (2.16). In fact, this is true if $v_{k+1} = d_k$; it is also true if $v_{k+1} = \tilde{d}_k$ and $d_k = \widehat{d}_k$, as projecting leaves into the ε_k-subdifferential. The only case where the property may not hold is when $d_k = \tilde{d}_k$ and $v_{k+1} = \widehat{d}_k$ (cf. Lemma 2.4), since then we know only that the *projected object* is a ε_k-subgradient, which does not mean that the unprojected one is such.
4. The development of the present paragraph, upon which all the following results hinge, does *not* extend to "heavy ball" subgradient methods. This is so despite the fact that $v_{k+1} = (x_k - x_{k+1})/\nu_k$ would at first seem a good candidate for a definition of the deflection term which may fit the present theory; for instance, it would satisfy (2.15). However, such a definition may violate the crucial (2.11), which geometrically means that for all $x \in X$, the angle between $x - x_k$ and $-v_{k+1}$ is not larger than the angle between $x - x_k$ and $-\tilde{d}_k$. Consider the simple example where $X = \mathbb{R}^2_+$, $f(x) = \langle(-1,1), x\rangle$, $x_1 = (1,1)$, $\nu_1 = 2$, and the oracle outputs $g_1 = (-1,1) = \nabla f(x_1)$ (i.e., $\sigma_1 = 0$); in this case, $g_1 = \tilde{d}_1 = \widehat{d}_1 = d_1$, $\widehat{x}_2 = (3,-1)$, and $x_2 = (3,0)$. It is then easy to choose $x \in X$ that violates (2.11); in fact, $v_2 = (x_1 - x_2)/2 = (-1, 1/2) \notin \partial f_X(x_1) = \{g_1\}$. While the momentum term may be an approximate subgradient for a tolerance larger than ε_k, there does not seem to be any way to compute that error a priori. Indeed, all the available analyses

of "heavy ball" subgradient methods treat the momentum term basically as "noise"; therefore, either its use worsens the convergence properties of the algorithm [27], or the momentum parameter is required to vanish [35]. However, the approach where $\bar{g}_k = g_k$ and $v_{k+1} = d_k = \widehat{d}_k$ might be expected to numerically behave quite similarly to a "heavy ball" one; in fact, especially when ν_k is "small" and X is a polyhedron, one can expect the effect of the projection in (1.2) to be "small," i.e., that $d_k = \widehat{d}_k$ and $x_{k+1} - x_k$ be almost collinear. In this sense, the deflected conditional methods analyzed in this paper may offer a provably convergent alternative to "heavy ball" ones which is likely to be similarly effective; numerical experiments will be required to verify whether this is actually the case.

A final remark must also be done on a possible occurrence which does not seem to have been explicitly considered in previous analyses about approximate subgradient methods. Let's start with an obvious observation.

Observation 2.7. Let $\sigma^* = \limsup_{k\to\infty} \sigma_k < +\infty$ be the *asymptotic maximum error* of the black box; if $f^* > -\infty$, then no subgradient method can be guaranteed to provide a solution with absolute error strictly smaller than σ^* for all possible choices of the starting point x_1.

Proof. It is consistent with the hypotheses that $\sigma_k \geq \sigma^*$ for all k (say, σ_k is nonincreasing with the iterations and σ^* is the limit), x_1 is chosen in such a way that $f(x_1) = f^* + \sigma^*$, and the oracle always returns $g_k = 0$ when called upon x_1. Therefore, $d_k = 0$ and $x_k = x_1$ for all k whatever the choices in (1.2), (1.5), and (1.6): the algorithm is unable to move away from x_1; thus, the obtained error is σ^*. ∎

Subgradient methods typically assume $d_k \neq 0$; this is not an issue in the "exact" case because in the (very unlikely) event that $g_k = 0$ the algorithm can be promptly terminated as x_k is optimal. Analogously, for an approximate method with *constant error* $\sigma_k = \sigma^*$ for all k, finding $g_k = 0$ allows us to conclude that x_k is σ^*-optimal and therefore, to terminate the algorithm because no better accuracy can be expected, as shown above. However, it may rather be the case that σ_k is *not* constant; this, for instance, makes particular sense in *Lagrangian relaxation* (cf. section 5.1). So, the case may arise where $\sigma_k > \sigma^*$ and $f^* + \sigma^* < f_k \leq f^* + \sigma^k$; the oracle may legally choose to return $g_k = 0$, which may result in $d_k = 0$. Stopping the algorithm is not a solution in this case, but something has to be done, because the stepsize formulae are not well defined; in practice, this would likely thrash any numerical code. Yet, the solution is simple: as shown in the above observation, $d_k = 0$ leaves no other choice to the algorithm than $x_{k+1} = x_k$. This is less dramatic than its loop-inducing aspect would initially suggest. In fact, if $\sigma_k = \sigma^*$, it is just an indication that the algorithm has converged to its maximum possible precision. If not, it means that the oracle can provide "more accurate" first-order information about the function in x_k "if instructed to do so"; in other words, calling the oracle again on the same x_k will either at length provide a nonzero g_k, or σ_k will converge to σ^*. We will just have to add specific provisions for handling the issue.

3. Stepsize-restricted approaches. We now proceed to proving convergence of the variants which choose to restrict the stepsize, leaving full scope for (almost) any choice of α_k. We will first present "abstract" conditions, in order to lay the foundations for fully implementable, target-level-like approaches.

3.1. Polyak stepsize. We start analyzing the—apparently new—*corrected Polyak stepsize*:

$$0 \leq \nu_k = \beta_k \frac{f_k - f^* - \gamma_k}{\|d_k\|^2} = \beta_k \frac{\lambda_k}{\|d_k\|^2}, \quad 0 \leq \beta_k \leq \alpha_k \leq 1, \tag{3.1}$$

where $\lambda_k = f_k - f^* - \gamma_k$. For (3.1) to have meaning, it is unavoidable to assume

$$f^* > -\infty. \tag{3.2}$$

Compared with the uncorrected Polyak stepsize (e.g., [20]), (3.1) is different in two aspects: first, $\beta_k \leq 1$ while usually $\beta_k \in (0, 2)$; second, the correction term "$-\gamma_k$" at the numerator. While the presence of γ_k may look surprising, we will show that a proper choice of the correction may substantially improve the convergence properties of the approach. However, it also introduces a significant complication in the analysis, as for $\gamma_k > 0$, λ_k can be negative; this leaves $\beta_k = 0 \Rightarrow \nu_k = 0$ as the only possible choice for satisfying (3.1). In other words, (3.1) implies

$$\lambda_k < 0 \ \Rightarrow \beta_k = 0. \tag{3.3}$$

Thus, whenever $\gamma_k > 0$ the algorithm can—unlike more common approaches—"visit" the same point more than once, meaning that special care is required to avoid stalling. This is also true for the case (that cannot be excluded; cf. Observation 2.7) when $d_k = 0$, which leaves (3.1) not well defined. In order to overcome this limitation, it is useful to introduce the following weaker form of (3.1):

$$0 \leq \nu_k \|d_k\|^2 \leq \beta_k \lambda_k, \quad 0 \leq \beta_k \leq \alpha_k \leq 1, \tag{3.4}$$

which is well defined even if $d_k = 0$. We will also impose the following strengthened form of (3.3):

$$\begin{array}{l} \lambda_k \geq 0 \ \Rightarrow \ (\alpha_k \ \geq) \, \beta_k \ \geq \ \beta^* > 0, \\ \lambda_k < 0 \ \Rightarrow \ \alpha_k = 0 \, (\Rightarrow \beta_k = 0); \end{array} \tag{3.5}$$

intuitively, this means that *positive* stepsizes do not vanish unless the (approximate) optimum is approached.

LEMMA 3.1. *Under* (3.1), (3.2), *and* (3.5), *it holds*

$$\begin{array}{l} \varepsilon_k \leq (1 - \alpha_k)(f_k - f^*) + \bar{\sigma}_k, \qquad \textit{where} \\ \bar{\sigma}_k = \left\{ \begin{array}{ll} \sigma_1 & k = 1 \\ (1 - \alpha_k)(\bar{\sigma}_{k-1}, -\alpha_{k-1}\gamma_{k-1}) + \alpha_k \sigma_k & \textit{otherwise.} \end{array} \right. \end{array} \tag{3.6}$$

Proof. Again, the proof is by induction on k. For $k = 1$, as in Lemma 2.6 $\alpha_1 = 1 \Rightarrow \varepsilon_1 = \sigma_1$. For the inductive step, using (2.16) we have

$$\begin{aligned} \varepsilon_k &= (1 - \alpha_k) \left(f_k - f_{k-1} - \langle v_k, x_k - x_{k-1} \rangle + \varepsilon_{k-1} \right) + \alpha_k \sigma_k \\ &\leq (1 - \alpha_k) \left(f_k - f_{k-1} + \nu_{k-1} \|d_{k-1}\|^2 + \varepsilon_{k-1} \right) + \alpha_k \sigma_k \end{aligned}$$

due to (2.15). Now, from (3.4) we have $\nu_{k-1}\|d_{k-1}\|^2 \leq \beta_{k-1}\lambda_{k-1}$, while (3.5) implies $\beta_{k-1}\lambda_{k-1} \leq \alpha_{k-1}\lambda_{k-1}$. In fact, this is true when $\lambda_{k-1} \geq 0$ since $\beta_{k-1} \leq \alpha_{k-1}$, and it is also true when $\lambda_{k-1} < 0$ since $\beta_{k-1} = \alpha_{k-1} = 0$. Hence, the inequality chain can be continued as

$$\begin{aligned} &(1 - \alpha_k) \left(f_k - f_{k-1} + \alpha_{k-1}(f_{k-1} - f^* - \gamma_{k-1}) + \varepsilon_{k-1} \right) + \alpha_k \sigma_k \\ &\leq (1 - \alpha_k)(f_k - f_{k-1} + \alpha_{k-1}(f_{k-1} - f^* - \gamma_{k-1}) \\ &\quad + (1 - \alpha_{k-1})(f_{k-1} - f^*) + \bar{\sigma}_{k-1}) + \alpha_k \sigma_k \\ &= (1 - \alpha_k)(f_k - f^*) + \left[\, (1 - \alpha_k) \, (\bar{\sigma}_{k-1} - \alpha_{k-1}\gamma_{k-1}) + \alpha_k \sigma_k \, \right], \end{aligned}$$

where in the second passage, we have invoked the inductive hypothesis. ∎

COROLLARY 3.2. *Under* (2.13), (3.1), *and* (3.2), *for each* $\bar{x} \in X$, *it holds*

$$\langle d_k, \bar{x} - x_k \rangle \leq \alpha_k (f^* - f_k) + [f(\bar{x}) - f^* + \bar{\sigma}_k]\,. \tag{3.7}$$

Proof. Using (2.12) (which holds due to (2.13) for Lemma 2.4), (2.16), and (3.6), one has

$$\langle d_k, \bar{x} - x_k \rangle \leq \langle v_{k+1}, \bar{x} - x_k \rangle \leq f(\bar{x}) - f_k + \varepsilon_k \leq f(\bar{x}) - f_k + (1-\alpha_k)(f_k - f^*) + \bar{\sigma}_k. \qquad \square$$

Note how much simpler the analysis is of the nondeflected case: no result like Lemma 3.1 is needed, as $\alpha_k = 1$ implies $\varepsilon_k = \sigma_k$ with no assumptions at all on the stepsize. The extra flexibility given by the added term γ_k allows us to obtain different estimates for the error at each iteration: it is easy to show by induction that in the interesting "extreme" case,

$$\gamma_k = \sigma_k \quad \Rightarrow \quad \bar{\sigma}_k = \alpha_k \sigma_k \tag{3.8}$$

deflection may even *increase* the accuracy of the available first-order information when optimality is "near" ($f_k \approx f^*$). Conversely, the *uncorrected* Polyak stepsize gives

$$\gamma_k = 0 \quad \Rightarrow \quad \bar{\sigma}_k = (1 - \alpha_k)\bar{\sigma}_{k-1} + \alpha_k \sigma_k. \tag{3.9}$$

Allowing to "aim at a different value than" f^* is necessary in practice, as f^* is usually unknown; besides, using a nonzero γ_k may be beneficial even if f^* is known, as discussed later on. Clearly, one would like a *nonnegative* γ_k: in fact, the error $\bar{\sigma}_k$ corresponding to a $\gamma_k < 0$ is always worse (not smaller) than the one corresponding to $\gamma_k = 0$. This means that while "aiming higher than f^*" ($\gamma_k > 0$) may be beneficial, "aiming lower than f^*" ($\gamma_k < 0$) is, in general, not. However, "aiming too high" is also dangerous, and $\gamma_k = \sigma_k$ is clearly the "extreme case," at least asymptotically; when $f^k - f^* \approx \sigma_k$, having $\gamma_k > \sigma_k$ systematically may make $\lambda_k < 0$ for all k, effectively grinding the algorithm to a halt (cf. Observation 2.7). So, a natural assumption would be $\gamma_k \leq \sigma_k$; this would immediately imply $\bar{\sigma}_k \geq 0$. However, this is *not* required to hold; rather, we rely on the following asymptotic results, which require much less stringent assumptions.

An important object in the analysis is $\bar{\sigma}^* = \limsup_{k \to \infty} \bar{\sigma}_k$. For the corrected Polyak stepsize with $\gamma_k = \sigma_k$, $\bar{\sigma}_k$ is "simple" (cf. (3.8)); thus, σ_k and $\bar{\sigma}_k$ "behave in the same way" for $k \to \infty$. We will show in the following that, due to (3.5), this is also true for the uncorrected Polyak stepsize ($\gamma_k = 0$). If γ_k can be negative, however, "extra noise" is added which depends upon

$$\bar{\gamma} \;=\; -\; \min\{\, \gamma^* = \liminf_{k \to \infty} \gamma_k \;,\; 0 \,\}\,. \tag{3.10}$$

LEMMA 3.3. *Under* (3.1), (3.2), *and* (3.5), *if* $\lambda_k \geq 0$ *for* infinitely many k, *then*

$$\bar{\sigma}^* \leq \sigma^* + \bar{\gamma}(1 - \beta^*)/\beta^*.$$

Proof. Note that whenever $\lambda_k < 0 \Rightarrow \alpha_k = 0$, all the information generated in iteration k is "lost" to subsequent iterations. In fact,

$$\begin{aligned}
\bar{\sigma}_{k+1} &= (1 - \alpha_{k+1})(\bar{\sigma}_k - \alpha_k \gamma_k) + \alpha_{k+1}\sigma_{k+1} \\
&= (1 - \alpha_{k+1})\,(\,(1 - \alpha_k)(\bar{\sigma}_{k-1} - \alpha_{k-1}\gamma_{k-1}) + \alpha_k \sigma_k - \alpha_k \gamma_k\,) + \alpha_{k+1}\sigma_{k+1} \\
&= (1 - \alpha_{k+1})(\bar{\sigma}_{k-1} - \alpha_{k-1}\gamma_{k-1}) + \alpha_{k+1}\sigma_{k+1},
\end{aligned}$$

and the same obviously also happens to all other relevant algorithmic quantities, e.g., x_k and d_k. Thus, assuming only that $\lambda_k \geq 0$ infinitely many times, we can restrict our attention to the iterations where this happens ($\Rightarrow \alpha_k \geq \beta^*$ due to (3.5)) and simply disregard all the others. Note that for $\gamma_k \leq 0$ (e.g., the uncorrected Polyak stepsize), the hypothesis $\lambda_k \geq 0$ is always verified.

We want to prove that for each $\varepsilon > 0$ and all sufficiently large h, one has $\bar{\sigma}_h \leq \sigma^* + \bar{\gamma}(1-\beta^*)/\beta^* + \varepsilon$. By the definition of σ^* and $\bar{\gamma}$, an analogous result holds for the "original" sequences: however, chosen a fixed constant $q > 0$, for a sufficiently large k, $\sigma_h \leq \sigma^* + \varepsilon/4$ and $-\gamma_h \leq \bar{\gamma} + q\varepsilon$ for all $h \geq k$. It is then easy to verify by induction that for $h \geq k$,

$$\bar{\sigma}_h \leq \bar{\sigma}_k(1-\beta^*)^{h-k} + \sigma^* + \bar{\gamma}(1-\beta^*)/\beta^* + \varepsilon/2. \tag{3.11}$$

In fact, the result is clearly true for $h = k$, while for the inductive step

$$\begin{aligned}\bar{\sigma}_h &= (1-\alpha_h)(\bar{\sigma}_{h-1} - \alpha_{h-1}\gamma_{h-1}) + \alpha_h\sigma_h \\ &\leq (1-\alpha_h)\left[\,\bar{\sigma}_k(1-\beta^*)^{h-1-k} + \sigma^* + \bar{\gamma}/\beta^* + q\varepsilon + \varepsilon/2\,\right] + \alpha_h(\sigma^* + \varepsilon/4) \\ &\leq (1-\beta^*)\bar{\sigma}_k(1-\beta^*)^{h-1-k} + \sigma^* + \bar{\gamma}(1-\beta^*)/\beta^* + \varepsilon[\,(1-\alpha_h)(q+1/2) + \alpha_h/4\,],\end{aligned}$$

where in the second step, we have used $-\alpha_{h-1}\gamma_{h-1} \leq \alpha_{h-1}(\bar{\gamma} + q\varepsilon) \leq \bar{\gamma} + q\varepsilon$ and the inductive hypothesis. Now, since $q + 1/2 > 1/4$, we have

$$(1-\alpha_h)(q+1/2) + \alpha_h/4 \leq (1-\beta^*)(q+1/2) + \beta^*/4 = (1-\beta^*)q - \beta^*/4 + 1/2.$$

Thus, by ensuring that $(1-\beta^*)q - \beta^*/4 \leq 0$ (e.g., choosing $q = \beta^*/(4(1-\beta^*))$), one has finally proven that (3.11) holds. It is now sufficient to choose $h \geq k$ such that $\bar{\sigma}_k(1-\beta^*)^{h-k} \leq \varepsilon/2$ to prove the thesis. □

While Lemma 3.3 provides a convenient estimate for the case where nothing can be said upon γ_k, it is clear that for γ_k "large enough" w.r.t. σ_k, something more can be said: in fact, for $\gamma_k = \sigma_k$, as we have already noted, $\bar{\sigma}_k = \alpha_k\sigma_k$ (without any assumption on λ_k). The "extra" factor α_k in the error estimate is relevant for the convergence analysis, as we shall see soon. However, replicating it when $\gamma_k < \sigma_k$ (but it is "large enough") is not straightforward; a useful result is the following.

LEMMA 3.4. *Under conditions* (3.1), (3.5), $\lambda_k \geq 0$ *for* infinitely many k *and*

$$\gamma^* \;\geq\; \xi\sigma^* \qquad\qquad \xi \in [0,1]: \tag{3.12}$$

- *for any $\varepsilon > 0$ there exists a k such that for all $h \geq k$,*

$$\gamma_k \;\geq\; \xi\sigma_h^k - \varepsilon; \tag{3.13}$$

- *for any $\varepsilon > 0$ there exists a k such that for all* large enough *$h \geq k$,*

$$\bar{\sigma}_h \;\leq\; \sigma_h^k(\,1 - (1-\alpha_h)\xi\,) + \varepsilon, \tag{3.14}$$

where $\sigma_h^k \;=\; \max\{\,\sigma_p \;:\; h \geq p \geq k\,\} \;\leq\; \sigma_\infty^k \;=\; \sup\{\,\sigma_p \;:\; p \geq k\,\}$.

Proof. As in Lemma 3.3, we can assume that $\lambda_k \geq 0 \Rightarrow \alpha_k \geq \beta^*$ at every iteration. From the definitions of γ^* and σ^*, however, fixed ε_1 and ε_2, for large enough k one has $\gamma_k \geq \inf_{p \geq k}\ \gamma_p \geq \gamma^* - \varepsilon_1$ and $\sigma_h^k \leq \sigma_\infty^k \leq \sigma^* + \varepsilon_2$ for all $h \geq k$, combining which gives (3.13). As for (3.14), choose any $\varepsilon > 0$ and select a sufficiently large k (which exists for (3.13)) such that

$$\gamma_k \;\geq\; \xi\sigma_h^k - \varepsilon\beta^*/(2-2\beta^*).$$

Then, we can prove by induction that for all $h \geq k$,

$$\bar{\sigma}_h \;\leq\; \bar{\sigma}_k(1-\beta^*)^{h-k} + \sigma_h^k(1-(1-\alpha_h)\xi) + \varepsilon/2. \tag{3.15}$$

In fact, the result is clearly true for $h = k$, while for the inductive step

$$\begin{aligned}
\bar{\sigma}_h &= (1-\alpha_h)(\bar{\sigma}_{h-1} - \alpha_{h-1}\gamma_{h-1}) + \alpha_h\sigma_h \\
&\leq (1-\alpha_h)\left[\, \bar{\sigma}_k(1-\beta^*)^{h-1-k} + \sigma_{h-1}^k(1-(1-\alpha_{h-1})\xi) + \varepsilon/2 - \alpha_{h-1}\gamma_{h-1} \,\right] + \alpha_h\sigma_h \\
&\leq \text{(induction)}\ (1-\beta^*)\bar{\sigma}_k(1-\beta^*)^{h-1-k} + (1-\alpha_h)\big[\, \sigma_{h-1}^k(1-(1-\alpha_{h-1})\xi) + \varepsilon/2 \\
&\quad \text{(choice of } k) - \alpha_{h-1}(\xi\sigma_{h-1}^k - \varepsilon\beta^*/(2-2\beta^*)) \,\big] + \alpha_h\sigma_h \\
&\leq [\beta^* \leq \alpha_h]\bar{\sigma}_k(1-\beta^*)^{h-k} + (1-\alpha_h)\left[\, \sigma_{h-1}^k(1-\xi) + \varepsilon/(2-2\beta^*) \,\right] + \alpha_h\sigma_h \\
&\leq [\alpha_{h-1} \leq 1]\bar{\sigma}_k(1-\beta^*)^{h-k} + (1-\alpha_h)\sigma_h^k(1-\xi) + \varepsilon/2 + \alpha_h\sigma_h^k \\
&= \bar{\sigma}_k(1-\beta^*)^{h-k} + \sigma_h^k(\, 1-(1-\alpha_h)\xi \,) + \varepsilon/2,
\end{aligned}$$

where in the penultimate line we have used $\sigma_h \leq \sigma_h^k$, $\sigma_{h-1}^k \leq \sigma_h^k$, $\beta^* \leq \alpha_h$, and $\xi \leq 1$. It is now sufficient to choose h large enough such that $\bar{\sigma}_k(1-\beta^*)^{h-k} \leq \varepsilon/2$ to prove (3.14). □

Hence, taking a "sufficiently large" γ_k asymptotically "shaves away" a fraction of $(1-\alpha_k)$, depending on ξ, from σ^*; for $\xi = 1$, one has $1-(1-\alpha_k)\xi = \alpha_k$ as expected. Note that the hypothesis "$\lambda_k \geq 0$ sufficiently often," crucial for both the lemmas above, is by no means trivial to attain for a *positive* γ_k.

Given the above results, convergence of the uncorrected Polyak stepsize under conditions (3.5) and (3.9) can be partly analyzed using results from [20]. In particular, for an *exact* oracle ($\sigma_k \equiv 0$), condition (3.1) turns out to imply [20, equation (7.28)], i.e.,

$$\text{there exist } \xi \in [0,1) \quad \varepsilon_k \leq \frac{1}{2}\xi(2-\beta_k)(f_k - f^*).$$

In fact, (3.1), and therefore, (3.6) hold; furthermore, since $\beta^* \leq \beta_k \leq \alpha_k$, one has

$$1-\alpha_k \leq 1-\beta_k \leq (1-\beta_k/2) - \beta_k/2 = (1-\beta_k/2) - \beta^*/2.$$

Thus, choosing ξ such that

$$1 > \xi \geq 1-\beta^*/(2-\beta_k) \geq 1-\beta^*/2 > 0,$$

our conditions imply [20, (7.28)]; under the additional assumption $X^* \neq \emptyset$, [20, Theorem 7.17(ii)] proves convergence to an optimal solution.

The same reference also allows us to (partly) analyze the case with error ($\sigma^* > 0$) of an "asymptotically nondeflected" method, i.e., one where $\lim_{k\to\infty}\alpha_k = 1$; this requires conditions ensuring that the sequence f_k is bounded (above), plenty of which are analyzed in [20, section 6]. In fact, in this case (3.6) and Lemma 3.3 imply that $\limsup_{k\to\infty}\varepsilon_k = \sigma^*$, and we can invoke [20, Theorem 7.17(i)] to conclude that

$$f^\infty \leq f^* + 2\sigma^*/(2-\beta_{\max}), \tag{3.16}$$

where $\beta_{\max} = \sup_k\ \beta_k (\leq 1)$. However, the above results are not completely satisfactory: the convergence of the case with errors is not established unless under a very strong condition, and (3.16) implies that the algorithm may only be able to find a solution whose accuracy is *twice as bad* as the inherent oracle error σ^*, unless the

maximum step is "artificially restricted," possibly impacting practical convergence rates (actually, $\beta_k \in (0,2)$ in [20], thus, that multiplying factor can become arbitrarily large). We, therefore, provide a specific analysis of the more general deflected conditional approximate method.

THEOREM 3.5. *Under conditions* (2.13), (3.1), (3.2), *and* (3.5), *it holds that*

(i) *let* $\Delta = \sigma^* + \bar{\gamma}(\ (1-\beta^*)/\beta^* + \alpha_{\max}/2\ )$, $\alpha_{\max} = \sup_k\ \alpha_k\ (\le 1)$, $\Gamma = \inf_k\ 2\alpha_k - \beta_k (\ge \beta^*)$; *if* $\limsup_{k\to\infty} \gamma_k \le 2\Delta/\Gamma$, *then*

$$f^\infty \le f^* + 2\Delta/\Gamma;$$

(ii) *if* $\gamma^* = \xi\sigma^*$ *for* $\xi \in [0,1]$ $(\Rightarrow \bar{\gamma} = 0)$, *then*

$$f^\infty \le f^* + \sigma^*(\ \xi + 2(1-\xi)/\Gamma\ );$$

(iii) *under choice* (3.8), $f^\infty \le f^* + \sigma^*$; *furthermore, if* $X^* \ne \emptyset$, *then a subsequence of* $\{x_k\}$ *converges to some* $x^\infty \in X$ *such that* $f(x^\infty) = f^\infty$, *and if, in addition,* $\sigma^* = 0$, *then the whole sequence converges to* x^∞, *and* $x^\infty \in X^*$.

Proof. For any $\bar{x} \in X$, using (2.9) one has

$$\|x_{k+1} - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 \le -2\nu_k \langle d_k, x_k - \bar{x}\rangle + \nu_k^2 \|d_k\|^2 = \kappa_k.$$

Due to (3.5), $\lambda_k < 0 \Rightarrow \nu_k = \alpha_k = \beta_k = 0 \Rightarrow \kappa_k = 0$; furthermore, $d_k = 0 \Rightarrow \kappa_k = 0$. In fact, the algorithm may "visit" the same point more than once, either because $\lambda_k < 0$ or because $d_k = 0$; in all other cases,

$$\begin{aligned}
\kappa_k &= -2\beta_k \frac{\lambda_k}{\|d_k\|^2}\langle d_k, x_k - \bar{x}\rangle + \beta_k^2 \frac{\lambda_k^2}{\|d_k\|^2} \\
&\le \text{ (for (3.1))}\ 2\beta_k \frac{\lambda_k}{\|d_k\|^2}(\alpha_k(f^* - f_k) + [f(\bar{x}) - f^* + \bar{\sigma}_k]) + \beta_k^2 \frac{\lambda_k^2}{\|d_k\|^2} \\
&= \text{ (for (3.7))}\ \beta_k \frac{\lambda_k}{\|d_k\|^2}\,[\ 2(\alpha_k(f^* - f_k) + f(\bar{x}) - f^* + \bar{\sigma}_k) + \beta_k\lambda_k\ ] \\
&= \beta_k \eta_k, \qquad (3.17)
\end{aligned}$$

where $\eta_k = -(2\alpha_k - \beta_k)(f_k - f^*) + 2(f(\bar{x}) - f^* + \bar{\sigma}_k) - \beta_k\gamma_k$.

Point i. Assume by contradiction that for some $\varepsilon > 0$, one has $f_k - f^* \ge f^\infty - f^* \ge 2\Delta/\Gamma + \varepsilon$. From the hypothesis on γ_k, it follows that at length $\lambda_k \ge \varepsilon/2$ (clearly, taking huge positive γ_k will easily force λ_k to be always zero, and this has to be avoided; however, this point is actually useful only when $\gamma_k < 0$). Then, $d_k = 0$ can only happen finitely many times. In fact, due to (2.13) we know that $d_k \in \partial_{\varepsilon_k} f(x_k)$ (cf. Remark 3 after Lemma 2.6); thus, $d_k = 0 \Rightarrow x_k$ is a ε_k-optimal solution, i.e., $f_k - f^* \le \varepsilon_k$. From (3.6), this gives $f_k - f^* \le \bar{\sigma}_k/\alpha_k$. So, assume this happens infinitely many times; from Lemma 3.3, for any $\varepsilon' > 0$ there exist some k such that

$$f_k - f^* \le (\sigma^* + \bar{\gamma}(1-\beta^*)/\beta^* + \varepsilon')/\alpha_k.$$

On the other hand, from the initial ab absurdo hypothesis, the definition of Δ and the fact that $\Gamma \le 2\alpha_k - \beta_k \le \alpha_k$ one has

$$f_k - f^* \ge 2\Delta/\Gamma + \varepsilon \ge (\sigma^* + \bar{\gamma}((1-\beta^*)/\beta^* + \alpha_{\max}/2))/\alpha_k + \varepsilon,$$

which easily yields a contradiction. Thus, at length $\lambda_k \ge 0$ and $d_k \ne 0$, we then have

$$\begin{aligned}
\eta_k &\le -\Gamma(f_k - f^*) + 2(f(\bar{x}) - f^*) + 2(\sigma^* + \bar{\gamma}(1-\beta^*)/\beta^* + \varepsilon_1) + \beta_k(\bar{\gamma} + \varepsilon_2) \\
&\le \text{(for Lemma 3.3 and the definition of } \bar{\gamma}) \\
&\quad -2\Delta - \Gamma\varepsilon + 2\varepsilon_3 + 2(\sigma^* + \bar{\gamma}((1-\beta^*)/\beta^* + \alpha_{\max}/2)) + 2\varepsilon_1 + \varepsilon_2,
\end{aligned}$$

where $\varepsilon_3 = f(\bar{x}) - f^*$ can be chosen arbitrarily small by properly choosing $\bar{x}$ and ε_1, ε_2 can be chosen arbitrarily small by taking k large enough. Thus, picking ε_1, ε_2, and ε_3 small enough one has that $\eta_k \leq -\Gamma\varepsilon/2 < 0$ for all k. Hence, (3.17) shows that for all k large enough,

$$\|x_{k+1} - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 \leq -\beta_k \frac{\Gamma \lambda_k \varepsilon}{2\|d_k\|^2} < 0.$$

This implies that the sequence $\|x_k - \bar{x}\|$ is nonincreasing, i.e., $\{x_k\}$ is bounded; finiteness of f thus implies that $\{f_k\}$ is bounded above, and therefore, by (3.6), $\{\varepsilon_k\}$ is bounded above by some $\bar{\varepsilon} < +\infty$. Since the image of a compact set in int $dom\, f = \mathbb{R}^n$ under the $\bar{\varepsilon}$-subdifferential mapping is compact [18, Proposition XI.4.1.2], $\|d_k\|$ is bounded above by some constant $D < +\infty$. Furthermore, β_k and λ_k are bounded below; thus, summing between k and $h \geq k$ gives

$$0 \leq \|x_{h+1} - \bar{x}\|^2 \leq \|x_k - \bar{x}\|^2 - (k-h)\beta^* \Gamma \varepsilon^2 / \left(4D^2\right),$$

which, for h large enough, yields the contradiction.

Point ii. Assume by contradiction that for some $\varepsilon > 0$, one has $f_k - f^* \geq f^\infty - f^* \geq \sigma^*(\, \xi + 2(1-\xi)/\Gamma\,) + \varepsilon$. Since $\gamma^* = \xi\sigma^*$ and $2(1-\xi)/\Gamma \geq 0$, at length $\lambda_k \geq \varepsilon/2$. Also, $d_k = 0$ can happen only finitely many times; in fact, reasoning as in the previous case, we have, using Lemma 3.4, that if $d_k = 0$ for k large enough, then

$$f_k - f^* \leq \left(\, \sigma_h^k(\, 1 - (1-\alpha_h)\xi\,) + \varepsilon' \,\right)/\alpha_k = \sigma_h^k(\, \xi + (1-\xi)/\alpha_k\,) + \varepsilon'/\alpha_k$$

for any arbitrary $\varepsilon' > 0$ and all $h \geq k$. But for an arbitrary $\varepsilon'' > 0$, $\sigma_h^k \leq \sigma^* + \varepsilon''$ (again, for k large enough and all $h \geq k$). On the other hand, from the initial ab absurdo hypothesis and $\Gamma \leq \alpha_k$, one has

$$f_k - f^* \geq \sigma^*(\, \xi + 2(1-\xi)/\alpha_k\,) + \varepsilon,$$

which gives the desired contradiction (α_k is bounded below). Then, we can assume $\lambda_k \geq 0$ and $d_k \neq 0$ and examine the "crucial" quantity η_k. For this case, using again Lemma 3.4 we have

$$\eta_k \leq -(2\alpha_h - \beta_h)(f_h - f^*) + 2\varepsilon_3 + 2(\sigma_h^k(\, 1 - (1-\alpha_h)\xi\,) + \varepsilon_2) - \beta_h\left(\xi\sigma_h^k - \varepsilon_1\right),$$

where $\varepsilon_3 = f(\bar{x}) - f^*$ can be chosen arbitrarily small by properly choosing $\bar{x}$ and ε_1, ε_2 can be chosen arbitrarily small by taking k large enough and $h \geq k$ large enough. Using again $\sigma_h^k \leq \sigma^* + \varepsilon''$ for an arbitrary $\varepsilon'' > 0$ (k large enough and all $h \geq k$), the inequality chain can then be continued as

$$\begin{aligned}\eta_k &\leq -(2\alpha_h - \beta_h)\left(f_h - f^* - \xi\sigma_h^k\right) + 2\sigma_h^k(1-\xi) + 2(\varepsilon_3 + \varepsilon_2 + \varepsilon_1) \\ &\leq -(2\alpha_h - \beta_h)(f_h - f^* - \xi\sigma^*) + 2\sigma^*(1-\xi) + \mu,\end{aligned}$$

where μ can be chosen small enough by properly adjusting k, h, and $\bar{x}$. Thus, using the hypothesis, one obtains that η_k can be made to be negative and nonvanishing, which yields the contradiction as in the previous case.

Point iii. Finally, consider choice (3.8) and assume by contradiction that $f_k - f^* \geq f^\infty - f^* \geq \sigma^* + \varepsilon$; hence, $\gamma_k = \sigma_k \leq \sigma^* + \varepsilon/2 \Rightarrow \lambda_k \geq \varepsilon/2 > 0$ for k large enough. Again, $d_k = 0$ cannot happen infinitely many times; in fact, for $d_k = 0$ we have $f_k - f^* \leq \bar{\sigma}_k/\alpha_k = \sigma_k$ (cf. (3.8)), which immediately yields a contradiction with

$f_k - f^* \geq \sigma^* + \varepsilon$ if k can be arbitrarily large. Then, again at length we have $\lambda_k \geq 0$ and $d_k \neq 0$, and η_k simplifies to

$$\eta_k = -(2\alpha_k - \beta_k)\lambda_k + 2(f(\bar{x}) - f^*) \leq -\Gamma\lambda_k + 2\varepsilon_3, \tag{3.18}$$

where $\varepsilon_3 = f(\bar{x}) - f^*$ can be chosen arbitrarily small; since both Γ and λ_k are bounded away from zero, this yields a contradiction as in the previous cases. Clearly, this point is a very simplified form of Point ii; however, the hypothesis on the relationship between γ_k and σ_k is weaker than (3.12) for $\xi = 1$ if the sequences are nonmonotone.

Assume now that $X^* \neq \emptyset$: choosing as $\bar{x}$ some $x^* \in X^*$ shows (cf. (3.18)) that $\|x_{k+1} - x^*\|$ is nonincreasing, and therefore, $\{x_k\}$ is bounded. Extracting a subsequence k_i such that $f_{k_i} \to f^\infty$ and a subsubsequence converging to some $x^\infty \in X$ (which exists by boundedness of $\{x_k\}$ and closedness of X), we have that $f(x^\infty) = f^\infty$ by continuity of f. Now, if $\sigma^* = 0$, then clearly $x^\infty \in X^*$; thus, $\liminf_{k\to+\infty} \|x_{k+1} - x^\infty\| = 0$. But we know that at length $\|x_{k+1} - x^\infty\|$ is nonincreasing, therefore, $\{x_k\} \to x^\infty$. □

The previous theorem not only generalizes [20, Theorem 7.17] to the deflected context, but also significantly improves on known results even for the nondeflected case. For the original Polyak stepsize (3.9), the extra "noise" introduced by deflection considerably worsens the error bound in the "inexact" case: only by asking that $\lim_{k\to\infty} \alpha_k = 1$—that is, asymptotically inhibiting deflection—*and* by requiring very short steps (which is presumably bad in practice), a method with an error close to σ^* is obtained. Besides, convergence of iterates seems to be lost for good. However, the corrected stepsize (3.8) makes up for both issues and even improves the convergence results: the final error is exactly σ^*, the minimum possible one (cf. Observation 2.7), and convergence of subsequences is attained also in the inexact case. If choosing the correction equal to the error ($\gamma_k = \sigma_k$) is not possible, e.g., because σ_k is unknown, ensuring that at least asymptotically the two are related helps in bounding the final error; this will be very useful in the next section.

It may be worth remarking that the algorithm may exhibit some "nonstandard" behavior:

- As already remarked (first in Lemma 3.3), it may happen that for some $k < h$, one has $\lambda_{k-1} > 0$, $\lambda_p \leq 0$ for all $k \leq p < h$, and $\lambda_h > 0$; that is, the algorithm has "got stuck" in x_{k-1} for some iterations, but it has finally "escaped" at iteration h. Then, it is clear that all the information generated at steps $k, \ldots, h-1$ is "lost": since $\alpha_p = 0$ for $k \leq p < h$, $d_{h-1} = d_k$ and $\varepsilon_{h-1} = \varepsilon_k$. Essentially, all the "useless" information has been discarded, and the algorithm has resumed business as usual as soon as a "useful" (approximate) subgradient has been found.
- It may happen that at some iteration k, the algorithm finds an x_k with an accuracy $f_k - f^*$ greater than, or equal to, that prescribed by Theorem 3.5; then, an infinitely long tail of iterations may start where $\lambda_h < 0 \Rightarrow \alpha_h = \beta_h = \nu_h = 0$ and/or $d_h = 0$, and therefore, the algorithm will "get stuck" on x_k ($x_h = x_k$) for all $h > k$. This is not an issue though: a solution with the "maximum possible" accuracy has, indeed, been *finitely* obtained, and, as in the proof of Observation 2.7, the subsequent iterations are only useful to "wait for σ_h to converge to σ^*."

In hindsight, (3.8)—and therefore, its "approximate version" (3.12)—also has a clear rationale. Whenever one obtains a function value that is off the optimal one by less than the current accuracy of the function computation (that is, $f_k \leq f^* + \sigma_k$), then

either the algorithm should be promptly terminated or the accuracy of the function computation should be increased (σ_k decreased). In fact, as shown in Observation 2.7, in such a situation $g_k = 0 \in \partial_{\sigma_k} f(x_k)$ may be legally returned by the black box (x_k is a σ_k-optimal solution); in other words, "basically any" g_k can be generated by the black box from that moment on, and no improvement in the function value should be expected, except by pure chance, unless σ_k eventually decreases enough to allow for positive stepsizes in (3.1). However, condition (3.8) does not come for free: it assumes *knowledge of* σ_k that is not required by (3.1)—which already is not implementable in general as f^* is unknown—with (3.9). Yet, both requirements can be done away simultaneously, disposing of (3.2) in the process, with the target value approaches described next.

3.2. Target value stepsize. The methods of section 3.1 are, in most situations, not directly implementable as they require knowledge of the value f^*, obtaining which is typically the main reason why (1.1) is solved in the first place. In the nondeflected setting, this can be avoided with the use of other forms of stepsizes, such as *diminishing/square summable* ones [20], that do not require knowledge of f^*. However, when deflecting, (some form of) knowledge of f^* is required anyway in order to bound the maximum error ε_k of d_k as a subgradient in x_k, which is central in our analysis; this makes diminishing stepsizes less attractive in our context.

There is another known workaround for this problem: a *target value* stepsize, whereby f^* is approximated by some estimate, that is properly revised as the algorithm proceeds. The usual form of the estimate is the *target level* $f^k_{lev} = f^k_{ref} - \delta_k$, where $f^k_{ref} \geq f^*$ is the *reference value* and $\delta_k > 0$ is the *threshold*. The target level is used instead of f^* in the stepsize formula (3.1), yielding

$$0 \leq \nu_k = \beta_k \frac{f_k - f^k_{lev}}{\|d_k\|^2}, \quad 0 \leq \beta_k \leq \alpha_k \leq 1. \tag{3.19}$$

While (3.19) looks to be an *uncorrected* target level stepsize, δ_k plays the role of γ_k here. In fact, we can exploit the general results of section 3.1 in this setting by simply noting that the "crucial" part of (3.19) is

$$\lambda_k = f_k - f^k_{lev} = f_k - f^* - \left(f^k_{ref} - f^* - \delta_k\right);$$

that is, the (uncorrected) level stepsize is a special case of the general corrected stepsize where the *actual correction*

$$\gamma_k = f^k_{ref} - f^* - \delta_k \tag{3.20}$$

is unknown. This also implies that target-level approaches do not require knowledge of σ_k.

A small technical hurdle need be addressed now. Let $f^k_{rec} = \min_{h \leq k} f(x_h)$ be the current *record value*, and $f^\infty_{rec} = \lim_{k \to \infty} f^k_{rec}$; if $f^\infty_{rec} = -\infty$, then $f^* = -\infty$ and the algorithm has clearly constructed an optimizing sequence. Whenever $f^\infty_{rec} > -\infty$, instead, we will need to invoke Theorem 3.5 in order to prove our accuracy estimates; however, just because $f^\infty_{rec} > -\infty$ we cannot assume that $f^* > -\infty$, and therefore, we have no guarantee that (3.20) is well defined. This is not really an issue, because it is possible to replace f^* in the analysis of the previous paragraph with a *feasible target* $\bar{f}$, i.e.,

$$\bar{f} > -\infty \quad , \quad \bar{f} \geq f^* \quad , \quad \bar{f} \leq f^\infty_{rec} \ \left(\Rightarrow f_k - \bar{f} \geq 0\right). \tag{3.21}$$

It is easy to verify that one can replace f^* with $\bar{f}$ in (3.1), so that it is well defined even if $f^* = -\infty$ and obtain that Lemma 3.1, Corollary 3.2, and Theorem 3.5 (but obviously the last part of Point iii) hold with $\bar{f}$ replacing f^* and without a need for assumption (3.2). This allows us to define

$$\gamma_k = f^k_{ref} - \bar{f} - \delta_k, \tag{3.22}$$

which is finite even if $f^* = -\infty$. The subsequent analysis then centers on this γ_k, i.e., on how well f^k_{ref} approximates $\bar{f}$ (f^*). This obviously depends on how f^k_{ref} and δ_k are updated; we will analyze both the strategies proposed in the literature.

3.2.1. Nonvanishing threshold. In this approach, f_{ref} is updated in the very straightforward way $f^k_{ref} = f^k_{rec}$ so that $f^\infty_{ref} = \lim_{k\to\infty} f^k_{ref} = f^\infty_{rec}$. An immediate consequence of this choice is that, since $\lambda_k = f_k - f^k_{rec} \geq 0$, one has $\lambda_k \geq \delta_k$. The threshold can also be updated in a very simple way: whenever "sufficient decrease" is detected, it can be reset to any "large number," otherwise, it has to be decreased, provided only that it does not vanish. The abstract property is

$$\text{either} \qquad f^\infty_{ref} = -\infty \qquad \text{or} \qquad \liminf_{k\to\infty} \delta_k = \delta^* > 0, \tag{3.23}$$

and one quite general (and simple) way of implementing it is

$$\delta_{k+1} \in \begin{cases} [\ \delta^*\ ,\ \infty\ ) & \text{if } f_{k+1} \leq f^k_{lev}, \\ [\ \delta^*\ ,\ \max\{\ \delta^*\ ,\ \mu\delta_k\ \}\ ] & \text{if } f_{k+1} > f^k_{lev}, \end{cases}$$

where $\mu \in [0, 1)$. Indeed, if $f_{k+1} \leq f^k_{lev}$ happens finitely many times, then $f^\infty_{ref} > -\infty$, and after the last time δ_k is nonincreasing and has limit δ^*. Otherwise, taking subsequences if necessary, at length $f^{k+1}_{ref} = f^{k+1}_{rec} \leq f_{k+1} \leq f^k_{ref} - \delta_k$, and δ_k is bounded away from zero; therefore, $f^\infty_{ref} = -\infty$ (which incidentally proves that f is unbounded below). For this approach, the following convergence result can be proven.

THEOREM 3.6. *Under conditions* (2.13) *and* (3.5), *the algorithm employing the level stepzise* (3.19) *with threshold condition* (3.23) *attains either* $f^\infty_{ref} = -\infty = f^*$ *or* $f^\infty_{ref} \leq f^* + \xi\sigma^* + \delta^*$, *where* $0 \leq \xi = \max\{\ 1 - \delta^*\Gamma/2\sigma^*\ ,\ 0\ \} < 1$.

Proof. If $f^\infty_{ref} = -\infty$, then clearly $f^* = -\infty$. Hence, assume $f^\infty_{ref} > -\infty$; from (3.23) we have $\delta^* > 0$, and furthermore,

$$\gamma^* = \liminf_{k\to\infty} \gamma_k = f^\infty_{ref} - \bar{f} - \delta^* \tag{3.24}$$

for any feasible target $\bar{f}$ (cf. (3.21)). Also, because $\lambda_k \geq \delta_k$, at length $\lambda_k \geq \delta^*/2 > 0$. We will now prove that $f^\infty_{ref} \leq \bar{f} + \xi\sigma^* + \delta^*$, which gives the desired result if f^* is finite by simply taking $\bar{f} = f^*$. However, such a result also proves that if $f^* = -\infty$, then $f^\infty_{ref} = -\infty$ (argue by contradiction and take $\bar{f}$ small enough), thus completing the proof. We need to distinguish two cases:

Case I: $\delta^* \leq 2\sigma^*/\Gamma$. This implies that $\xi \geq 0$, and therefore, $\delta^* = 2(1-\xi)\sigma^*/\Gamma$. Assume by contradiction that $f^\infty_{ref} - \bar{f} > \xi\sigma^* + \delta^*$; from (3.24) this gives (3.12). Furthermore, since $f^\infty \geq f^\infty_{ref}$, we also obtain $f^\infty - \bar{f} > \xi\sigma^* + \delta^*$. Now, mirroring the proof of Theorem 3.5(ii) we can obtain

$$f^\infty - \bar{f} \leq \sigma^*(\ \xi + 2(1-\xi)/\Gamma\ ) = \xi\sigma^* + \delta^*$$

and therefore, a contradiction. The only difference in the proof, apart from $\bar{f}$ replacing f^*, is that γ^* can be *strictly larger* than $\xi\sigma^*$, and even larger than σ^*. This is not an

issue for Lemma 3.4 and therefore, for most of the proof; in particular, it is still true that η_k is negative and nonvanishing. The only reason why in Theorem 3.5(ii) one cannot directly assume $\gamma^* \geq \xi\sigma^*$ is that arbitrarily large γ_k may easily make $\lambda_k < 0$; thus, one needs a condition which, at length, ensures that λ_k is larger than a fixed positive threshold. However, here this is guaranteed by the fact that $\delta^* > 0$, so the proof readily extends.

Case II: $\delta^* > 2\sigma^*/\Gamma$. This implies $\xi = 0$. It is immediate to prove that $\gamma^* = f^\infty_{ref} - \bar{f} - \delta^* < 0$, which provides the expected estimate. In fact, assume by contradiction that $\gamma^* \geq 0$; this gives $\bar{\gamma} = -\min\{\gamma^*, 0\} = 0$, and therefore, by Theorem 3.5(i), $f^\infty - \bar{f} \leq 2\sigma^*/\Gamma < \delta^*$. On the other hand, $f^\infty \geq f^\infty_{ref}$ and $\gamma^* = f^\infty_{ref} - \bar{f} - \delta^* \geq 0$ give $f^\infty - \bar{f} \geq \delta^*$, a contradiction. ∎

For the finite case ($f^* > -\infty$), the above estimate compares favorably, when $\sigma^* > 0$, with that of [20, Theorem 7.19] (with "abstract" conditions on ε_k), which is $\sigma^* + \delta^*$. Actually, the term "$\xi\sigma^*$" in the estimate may look somewhat surprising; as $\xi < 1$, it may appear that choosing a "large" δ^* could help in reducing the final error. This clearly isn't so: $\xi\sigma^* + \delta^* = \sigma^*(\, \xi + 2(1-\xi)/\Gamma\,) \geq \sigma^*$ as $\Gamma \leq 1$.

3.2.2. Vanishing threshold. A vanishing $\{\delta_k\}$ sequence cannot be used in the proof of Theorem 3.6 because $\delta_{\min} > 0$ is used to ensure that λ_k does not vanish; this, however, introduces a source of error that worsens the convergence. If δ_k is to be allowed to vanish, then the same may happen to λ_k; however, for the argument to work it is actually only necessary that

$$\sum_{k=1}^{\infty} \lambda_k / \|d_k\|^2 = \infty. \tag{3.25}$$

In fact, in (3.17) one has

$$\|x_{k+1} - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 \leq \beta_k \lambda_k \eta_k / \|d_k\|^2,$$

where η_k can be bounded above by a negative quantity and β_k is bounded away from zero. This leaves the possibility open to vanishing δ_k, provided only that λ_k does not vanish "too quickly." Thus, the abstract property required instead of (3.23) is

$$\text{either} \quad f^\infty_{ref} = f^* = -\infty \quad \text{or} \quad \liminf_{k\to\infty} \delta_k = 0 \text{ and } (3.25)\,. \tag{3.26}$$

THEOREM 3.7. *Under conditions* (2.13) *and* (3.5), *the algorithm employing the level stepzise* (3.19) *with threshold condition* (3.26) *attains either* $f^\infty_{ref} = -\infty = f^*$ *or* $f^\infty_{ref} \leq f^* + \sigma^*$.

Proof. If $f^\infty_{ref} = -\infty$, there is nothing to prove; hence, assume $f^\infty_{ref} > -\infty$; from (3.22) and (3.26)

$$\gamma^* = \liminf_{k\to\infty} \gamma_k = f^\infty_{ref} - \bar{f} - \delta^* = f^\infty_{ref} - \bar{f} \tag{3.27}$$

for any feasible target $\bar{f}$ (cf. (3.21)). Now, assume by contradiction that $\gamma^* = f^\infty_{ref} - \bar{f} > \sigma^*$. Proceeding as in the proof of Theorem 3.5(ii) (with $\xi = 1$ and $\bar{f}$ replacing f^*), one obtains that η_k is negative and nonvanishing; using (3.25) and (3.5), one finally obtains $f^\infty \leq \bar{f} + \sigma^*$. On the other hand, $f^\infty \geq f^\infty_{ref}$ gives $f^\infty - \bar{f} > \delta^*$, a contradiction. As in the nonvanishing case, the above result finally implies that $f^* = -\infty \Rightarrow f^\infty_{ref} = -\infty$. ∎

Of course, we still have to clarify how the abstract condition (3.26) can be obtained. Fortunately, in the proof of Theorem 3.7 (Theorem 3.5), $\|d_k\|$ is bounded above by some $D < +\infty$ ($\{x_k\}$ is compact), so one can obtain the same result by asking that the series of

$$s_k = \|\widehat{x}_{k+1} - x_k\| = \nu_k \|d_k\| = \beta_k \lambda_k / \|d_k\| \tag{3.28}$$

(cf. (1.2) and (3.19)) diverges. This is much easier to attain, since s_k can be readily computed and managed.

LEMMA 3.8. *Under condition* (3.5), *the update strategy*

- $f^1_{ref} = f(x_1)$, $\delta_1 \in (0, \infty)$, $r_1 = 0$;
- *if* $f_k \le f^k_{ref} - \delta_k/2$ (sufficient descent condition), *then* $f^k_{ref} = f^k_{rec}$, $r_k = 0$;
- *else, if* $r_k > R$ (target infeasibility condition), *then* $\delta_k = \mu \delta_{k-1}$, $r_k = 0$;
- *otherwise,* $f^k_{ref} = f^{k-1}_{ref}$, $\delta_k = \delta_{k-1}$, $r_k = r_{k-1} + \|\widehat{x}_{k+1} - x_k\|$,

where $R > 0$ *and* $\mu \in (0, 1)$ *are fixed, attains either* $f^\infty_{ref} = -\infty$ *or* $\delta_k \to 0$ *and* $\sum_{k=1}^\infty s_k = \infty$.

Proof. If $f^\infty_{ref} = -\infty$, there is nothing to prove, so assume $f^\infty_{ref} > -\infty$. We first prove that the number of *resets*, i.e., the number of times in which r_k is set to zero (by either condition), is infinite. In fact, assume the contrary holds; for some k and all $h \ge k$, one would have

$$f^h_{ref} = f^k_{ref}, \quad \delta_h = \delta_k, \quad r_{h+1} = r_h + \|\widehat{x}_{h+1} - x_h\| \le R,$$

which implies $\|\widehat{x}_{h+1} - x_h\| \to 0$ and $\{x_k\}$ bounded. Therefore, as in the proof of Theorem 3.5 we have that $\|d_k\|$ is bounded above by some $D < +\infty$. But

$$\|\widehat{x}_{h+1} - x_h\| = \nu_h \|d_h\| = \beta_h \frac{f(x_h) - f^h_{lev}}{\|d_h\|} \ge \frac{\beta^*}{D} \left(f(x_h) - f^h_{lev} \right)$$

and therefore, at length $f(x_h) - f^h_{lev} = f(x_h) - f^h_{ref} + \delta_h \to 0$. Thus, at length $f(x_h) < f^h_{ref} - \delta_h/2$, a contradiction.

We can now prove that, out of the infinitely many resets, those due to the target infeasibility condition are infinitely many. In fact, if not then for some k and all $h \ge k$, one would have $\delta_h = \delta_k$, and since infinitely many resets due to the sufficient descent condition with nonvanishing δ_k are performed, then $f^\infty_{ref} = -\infty$, a contradiction. Thus, $\delta_k \to 0$ and $\sum_{k=1}^\infty s_k = \infty$. ∎

It would clearly be possible to allow for *increases* of δ_k in the above scheme, provided that this only happens finitely many times. Thus, the target management in Lemma 3.8 attains the convergence result of Theorem 3.7. The above treatment extends those of [33, 34, 26] for deflected target value approaches, which require analogous conditions to those of the present algorithm (in particular, $\alpha_k \ge \beta_k$), to considering inexact computation of the function and projection of directions.

4. Deflection-restricted approaches. We now proceed with the other main class of stepsize rules, i.e., *diminishing/square summable*:

$$\sum_{k=1}^\infty \nu_k = \infty, \quad \sum_{k=1}^\infty \nu_k^2 < \infty. \tag{4.1}$$

As previously mentioned, in our context these rules lose a part of their original appeal, w.r.t. Polyak-type rules, because in order to ensure convergence in the deflected case,

some conditions on the α_k multipliers have to be enforced which depend on the optimal value f^* (cf. Figure 2.1). However, as in the stepsize-restricted case, approximations to f^* can be used; hence, also in this case we will first analyze abstract rules, in order to later move to implementable ones.

4.1. Abstract deflection condition. Our analysis centers on the *deflection condition*

$$0 \leq \zeta_k = \frac{\nu_{k-1}\|d_{k-1}\|^2}{(\ f_k - f^* - \gamma_k\ ) + \nu_{k-1}\|d_{k-1}\|^2} \leq \alpha_k \leq 1, \tag{4.2}$$

where, as usual, we assume $\alpha_1 = 1 \Rightarrow d_1 = g_1$; clearly, (3.2) is required for the condition to have meaning. We call (4.2) a *corrected* deflection condition due to the presence of the parameter γ_k in $\lambda_k = f_k - f^* - \gamma_k$ at the denominator. As we will see, γ_k plays a very similar role as in (3.1): its "optimal" value is σ_k, and the farthest it is from this optimal choice, the worst are the convergence properties of the algorithm. In particular, the simplest choice $\gamma_k = 0$ gives the *uncorrected* deflection condition; an interesting property of this choice is that, unless an optimal solution is finitely attained (that is, $f_k = f^*$ for some k), it implies $\zeta_k < 1$ for all k; therefore, some amount of deflection is possible at *every iteration.*

This is a fortiori true if $\gamma_k < 0$, which, as in the stepsize-restricted case, turns out to be a bad choice; unfortunately, the better choice $\gamma_k > 0$ gives rise to the possibility that ζ_k is undefined. To avoid this, we first rewrite (4.2) as

$$\nu_{k-1}\|d_{k-1}\|^2 \leq \alpha_k \left(\lambda_k + \nu_{k-1}\|d_{k-1}\|^2\right) \tag{4.3}$$

and then introduce the following condition, analogous to (3.5):

$$\begin{aligned} \lambda_k \geq 0 \ &\Rightarrow\ \alpha_k \geq \alpha^* > 0, \\ \lambda_k < 0 \ &\Rightarrow\ \alpha_k = 0\ (\Rightarrow \nu_k = 0). \end{aligned} \tag{4.4}$$

If $\lambda_k < 0$, the only way in which (4.3) can be satisfied is by forcing $\alpha_k = 0$, which, as in the stepsize-restricted case, also ensures that all information computed during such a "bad step" is disregarded. This in turn forces $\nu_k = 0$, unless in the special case where $d_k = 0$; both conditions, however, result in $\zeta_k = 0$. This mechanism automatically takes care, when $\gamma_k = \sigma_k$, of the possibility that $f_k < f^* + \sigma_k$, which has already been extensively commented upon; note that, unlike the stepsize-resticted case, $d_k = 0$ pose no specific challenge here. Of course, forcing $\nu_k = 0$ looks pretty much at odds with (4.1); in particular, (4.1) cannot possibly be satisfied if $\lambda_h < 0$ for all h larger than one given k, as it can be the case, e.g., if x_k is a σ^*-optimal solution (cf. Observation 2.7). Thus, the convergence analysis will have to take care of the above case ("optimal" solution found in a finite number of steps) separately.

We now start analyzing the properties of the iterates of a deflection-restricted approach employing the deflection condition (4.2) under assumption (4.4); the first three results closely mirror Lemma 3.1, Corollary 3.2, and Lemma 3.3, respectively.

LEMMA 4.1. *Under* (3.2), (4.2), *and* (4.4), *it holds*

$$\begin{aligned} &\varepsilon_k \leq f_k - f^* + \bar{\sigma}_k, \qquad \textit{where} \\ &\bar{\sigma}_k = \begin{cases} \sigma_1 - \gamma_1 & \textit{if } k = 1, \\ \alpha_k(\sigma_k - \gamma_k) + (1 - \alpha_k)\bar{\sigma}_{k-1} & \textit{otherwise.} \end{cases} \end{aligned} \tag{4.5}$$

Proof. As in the stepsize-restricted case, we can assume $\lambda_k \geq 0$; in fact, if this does not hold, then $\alpha_k = 0$ and "no trace" of the information generated at the kth

iteration remains, so we can restrict our attention only on the iterates k for which the property holds. Hence, we can proceed by induction: for $k = 1$, $\varepsilon_1 = \sigma_1 \leq \lambda_1 + \sigma_1 = f(x_1) - f^* + \sigma_1 - \gamma_1$. Then,

$$\begin{aligned}
\varepsilon_k &= (1-\alpha_k)\left(f_k - f_{k-1} - \langle v_k, x_k - x_{k-1}\rangle + \varepsilon_{k-1} \right) + \alpha_k\sigma_k \\
&\leq ((2.16))\ (1-\alpha_k)\left(f_k - f_{k-1} + \nu_{k-1}\|d_{k-1}\|^2 + \varepsilon_{k-1} \right) + \alpha_k\sigma_k \\
&\leq \text{ (by (2.15))}\ (1-\alpha_k)\left(f_k - f_{k-1} + \nu_{k-1}\|d_{k-1}\|^2 + f_{k-1} - f^* + \bar\sigma_{k-1} \right) + \alpha_k\sigma_k \\
&= \text{ (induction)}\ (1-\alpha_k)\left(f_k - f^* - \gamma_k + \nu_{k-1}\|d_{k-1}\|^2 \right) \\
&\quad + (1-\alpha_k)\bar\sigma_{k-1} + (1-\alpha_k)\gamma_k + \alpha_k\sigma_k \\
&\leq f_k - f^* - \gamma_k + (1-\alpha_k)\bar\sigma_{k-1} + (1-\alpha_k)\gamma_k + \alpha_k\sigma_k \\
&= \text{(by (4.2))}\ f_k - f^* + \bar\sigma_k. \qquad \square
\end{aligned}$$

The lemma confirms that $\gamma_k = \sigma_k$ is the best possible correction, as it minimizes the estimate of ε_k in (4.5); indeed, in that case one has $\bar\sigma_k = 0$ for all k. Of course, $\gamma_k > \sigma_k$ would be even better, except that it would not be possible to ensure that $\lambda_k \geq 0$ "often enough." Indeed, $\gamma_k \leq \sigma_k$ would imply $\bar\sigma_k \geq 0$, which is not true (and not required) in general.

COROLLARY 4.2. *Under* (3.2), (4.2), *and* (4.4), *for each* $\bar x \in X$ *it holds*

$$\langle v_{k+1}, \bar x - x_k\rangle \leq f(\bar x) - f^* + \bar\sigma_k. \tag{4.6}$$

Proof. Using (4.5), one has $\langle v_{k+1}, \bar x - x_k\rangle \leq f(\bar x) - f_k + \varepsilon_k \leq f(\bar x) - f^* + \bar\sigma_k$. $\square$

As in the stepsize-restricted case, $\bar\sigma_k$ "behaves as σ_k" for $k \to \infty$, unless $\gamma_k < 0$ in which case some "noise" appears.

LEMMA 4.3. *Under* (4.2) *and* (4.4), *if* $\lambda_k \geq 0$ *for* infinitely many k, *then*

$$\bar\sigma^* = \limsup_{k\to\infty} \bar\sigma_k \leq \sigma^* + \bar\gamma, \tag{4.7}$$

where $\bar\gamma$ *is defined as in* (3.10).

Proof. First, since $\lambda_k \geq 0$ for infinitely many k, we can disregard all iterations where it does not happen; due to (4.4), nothing actually happens in these. From the definition of σ^* and $\bar\gamma$, there exist some h such that, for all $k \geq h$,

$$\sigma_k \leq \sigma^* + \varepsilon \qquad \text{and} \qquad -\gamma_k \leq \bar\gamma + \varepsilon.$$

One can then prove by induction that for all $k \geq h$ it holds

$$\bar\sigma_k \leq (1-\alpha^*)^{k-h}\bar\sigma_h + \sigma^* + \bar\gamma + 2\varepsilon.$$

In fact, the statement is obviously true for $k = h$; for the general case we have

$$\begin{aligned}
\bar\sigma_k &= \alpha_k(\sigma_k - \gamma_k) + (1-\alpha_k)\bar\sigma_{k-1} \\
&\leq \alpha_k(\,\sigma^* + \bar\gamma + 2\varepsilon\,) + (1-\alpha_k)\left((1-\alpha^*)^{k-h-1}\bar\sigma_h + \sigma^* + \bar\gamma + 2\varepsilon \right) \\
&= \text{(induction)}\ (1-\alpha_k)(1-\alpha^*)^{k-h-1}\bar\sigma_h + \sigma^* + \bar\gamma + 2\varepsilon \leq (1-\alpha^*)^{k-h}\bar\sigma_h \\
&\quad + \sigma^* + \bar\gamma + 2\varepsilon.
\end{aligned}$$

The thesis now easily follows by taking h large enough. $\square$

As in the stepsize-restricted case, between the "optimal" correction (3.8) and the "bad one," $\gamma_k < 0$ one has a whole range of intermediate options, where γ_k is positive and "not too small" w.r.t. σ_k; this may take the form

$$\text{for all } k \text{ large enough}, \quad \gamma_k \geq \xi \sigma_k \qquad \xi \in [0, 1]. \tag{4.8}$$

To study this case, we consider the following handy sequence:

$$s_k = \alpha_k \sigma_k + (1 - \alpha_k) s_{k-1}.$$

LEMMA 4.4. *Under* (4.8), $\bar{\sigma}^* \leq (1 - \xi) \limsup_{k \to \infty} s_k$.

Proof. Let k be the index such that $\gamma_h \geq \xi \sigma_h$ for all $h > k$; we can prove by induction that

$$\bar{\sigma}_h \leq (1 - \xi) s_h + (1 - \alpha^*)^{h-k} (\bar{\sigma}_k - (1 - \xi) s_k)$$

for all $h \geq k$. The case $h = k$ is trivial; for the general inductive step,

$$\begin{aligned}
\bar{\sigma}_h &= \alpha_h (\sigma_h - \gamma_h) + (1 - \alpha_h) \bar{\sigma}_{h-1} \\
&\leq \alpha_h (1 - \xi) \sigma_h + (1 - \alpha_h) \left((1 - \xi) s_{h-1} + (1 - \alpha^*)^{h-1-k} (\bar{\sigma}_k - (1 - \xi) s_k) \right) \\
&\leq (1 - \xi)(\alpha_h \sigma_h + (1 - \alpha_h) s_{h-1}) + (1 - \alpha^*)^{h-k} (\bar{\sigma}_k - (1 - \xi) s_k) \\
&\leq (1 - \xi) s_h + (1 - \alpha^*)^{h-k} (\bar{\sigma}_k - (1 - \xi) s_k).
\end{aligned}$$

The result then follows taking the lim sup on both sides for $h \to \infty$. ∎

The $\{s_k\}$ sequence actually coincides with $\bar{\sigma}_k$ for the special choice $\gamma_k = 0$; thus, from Lemma 4.3 one immediately gets $\limsup_{k \to \infty} s_k \leq \sigma^*$, which therefore, leads, under (4.8), to a strengthened form of (4.7):

$$\bar{\sigma}^* \leq (1 - \xi) \limsup_{k \to \infty} s_k \leq (1 - \xi) \sigma^*. \tag{4.9}$$

Thus, as in the stespize-restricted case, we have three different settings concerning the "asymptotic accuracy"

$$\limsup_{k \to \infty} \varepsilon_k \leq \limsup_{k \to \infty} f_k - f^* + \bar{\sigma}_k$$

(cf. (4.5)) of the direction: other than from the error $f_k - f^*$, it depends on a term that is decreasing "the more γ_k is similar to σ_k." In particular, the term is $\sigma^* + \bar{\gamma}$ if nothing can be said on γ_k, it is $(1 - \xi)\sigma^*$ if (4.8) holds, and it is 0 if $\gamma_k = \sigma_k$ ($\xi = 1$).

We can now prove convergence of the approach; differently from the stepsize-restricted case, here we will need a weak form of boundedness of the iterates

$$D = \sup_k \|d_k\| < \infty. \tag{4.10}$$

This is true at the very least if X is compact and f finite everywhere, or if f is polyhedral and ε_k bounded by above. A number of *bounding strategies* are available to enforce this kind of (or even stronger) boundedness properties; see, e.g., [20, section 6] for a thorough discussion.

THEOREM 4.5. *If conditions* (3.2), (4.1), (4.2), (4.4), *and* (4.10) *hold, then*
(i) *let* $\gamma^{\sup} = \limsup_{k\to\infty} \gamma_k$*; then* $f^\infty \le f^* + \gamma^{\sup} + (\sigma^* + \bar\gamma)/\alpha^*$*;*
(ii) *under* (4.8), $f^\infty \le f^* + \sigma^*(\, 1 + (1-\xi)(1-\alpha^*)/\alpha^*\,)$*;*
(iii) *under choice* (3.8), $f^\infty \le f^* + \sigma^*$*; furthermore, if* $X^* \ne \emptyset$ *and* (2.13) *holds, then the sequence* $\{x_k\}$ *is convergent to some* x^∞ *such that* $f(x^\infty) = f^\infty$.

Proof. As previously anticipated, we must first do away with the finite termination case, i.e., that where the prescribed accuracy bounds are *finitely* attained at some iteration k. However, in this case there is nothing left to prove, so we must now argue by contradiction against the case where the bounds are not attained, even in the limit; for case (i), for instance, this means

$$f_k \ge f^* + \gamma^{\sup} + (\sigma^* + \bar\gamma)/\alpha^* + \varepsilon$$

for all k. It is then immediate to show that at length $\lambda_k > 0$. Thus, (4.4) and (4.1) are no longer at odds, and the above results (e.g., Lemma 4.3) can be safely invoked; actually, as in the previous cases we can restrict ourselves to the (infinite) subsequence where $\lambda_k > 0$ and disregard all other iterations. The argument for cases (ii) and (iii) is analogous.

For any $\bar x \in X$, from (2.9) one has

$$\|x_{k+1} - \bar x\|^2 - \|x_k - \bar x\|^2 \le -2\nu_k \langle x_k - \bar x, d_k\rangle + \nu_k^2 \|d_k\|^2.$$

Fixing any k and $h > k$, by summation we then have

$$\|x_h - \bar x\|^2 - \|x_k - \bar x\|^2 \le -2\sum_{j=k}^{h-1} \nu_j \langle x_j - \bar x, d_j\rangle + \sum_{j=k}^{h-1} \nu_j^2 \|d_j\|^2. \tag{4.11}$$

Hence, $l = \liminf_{k\to\infty} \langle d_k, x_k - \bar x\rangle \le 0$. In fact, assume by contradiction $\langle d_k, x_k - \bar x\rangle \ge \varepsilon > 0$ for all k; then, from (4.11) we get $\|x_h - \bar x\|^2 \to -\infty$ as $h \to \infty$. Now, let $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ be such that $(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) = \alpha^*\varepsilon/2$. Because $l \le 0$, there exists a subsequence $\{x_{k_i}\}$ such that

$$\langle d_{k_i}, x_{k_i} - \bar x\rangle \le \varepsilon_1 \tag{4.12}$$

(this is obvious if $l = -\infty$, otherwise, take a subsequence converging to l). Furthermore, from (2.15) and (4.10),

$$\langle v_{k+1}, x_k - x_{k+1}\rangle \le \nu_k \|d_k\|^2 \le D^2 \nu_k \to 0;$$

therefore, for large enough k,

$$\langle v_{k+1}, x_k - x_{k+1}\rangle \le \varepsilon_2. \tag{4.13}$$

Finally, we can choose $\bar x$ so that $f(\bar x) \le f^* + \varepsilon_3$. Then, taking i large enough for

(4.12) and such that k_i is large enough for (4.13), we have

$$\begin{aligned}
0 &\le \langle d_{k_i}, \bar{x} - x_{k_i}\rangle + \varepsilon_1 \\
&\le (\text{for (4.12)})\ \langle \tilde{d}_{k_i}, \bar{x} - x_{k_i}\rangle + \varepsilon_1 \\
&= (\text{for (2.10)})\ \alpha_{k_i}\langle \bar{g}_{k_i}, \bar{x} - x_{k_i}\rangle + (1-\alpha_{k_i})\langle v_{k_i}, \bar{x} - x_{k_i}\rangle + \varepsilon_1 \\
&\le (\text{for (1.6)})\ \alpha_{k_i}(f(\bar{x}) - f(x_{k_i}) + \sigma_{k_i}) + (1-\alpha_{k_i})\langle v_{k_i}, \bar{x} - x_{k_i}\rangle + \varepsilon_1 \\
&= [\bar{g}_{k_i} \in \partial_{\sigma_{k_i}} f_X(x_{k_i})]\alpha_{k_i}(f(\bar{x}) - f(x_{k_i}) + \sigma_{k_i}) + (1-\alpha_{k_i})\langle v_{k_i}, \bar{x} - x_{k_i-1}\rangle \\
&\quad + (1-\alpha_{k_i})\langle v_{k_i}, x_{k_i-1} - x_{k_i}\rangle + \varepsilon_1 \\
&\le \alpha_{k_i}(f(\bar{x}) - f(x_{k_i}) + \sigma_{k_i}) + (1-\alpha_{k_i})(f(\bar{x}) - f^* + \bar{\sigma}_{k_i-1}) \\
&\quad + (\text{for (4.6)})\ (1-\alpha_{k_i})\varepsilon_2 + \varepsilon_1 \\
&\le (\text{for (4.13)})\ f(\bar{x}) - \alpha_{k_i} f(x_{k_i}) + \alpha_{k_i}(\gamma_{k_i} + \sigma_{k_i} - \gamma_{k_i}) \\
&\quad + [\pm\gamma_{k_i}](1-\alpha_{k_i})\bar{\sigma}_{k_i-1} - (1-\alpha_{k_i})f^* + \varepsilon_2 + \varepsilon_1 \\
&\le [\alpha_{k_i} \ge 0] f^* - \alpha_{k_i} f(x_{k_i}) + \alpha_{k_i}\gamma_{k_i} + \bar{\sigma}_{k_i} - (1-\alpha_{k_i})f^* + \varepsilon_3 + \varepsilon_2 + \varepsilon_1 \\
&\le (\text{for (4.5)})\ \alpha_{k_i}(f^* - f(x_{k_i}) + \gamma_{k_i}) + \bar{\sigma}_{k_i} + \alpha^*\varepsilon/2.
\end{aligned}$$

Therefore, for large enough i,

$$f(x_{k_i}) \le f^* + \gamma_{k_i} + (\bar{\sigma}_{k_i} + \alpha^*\varepsilon/2)/\alpha_{k_i} \le f^* + \gamma_{k_i} + \bar{\sigma}_{k_i}/\alpha^* + \varepsilon/2. \tag{4.14}$$

Point i. The desired contradiction comes immediately from (4.14) by choosing i large enough so that $\gamma_{k_i} + \bar{\sigma}_{k_i}/\alpha^* \le \gamma^{\sup} + \bar{\sigma}^*/\alpha^* + \varepsilon/2$; thus, the theorem is proved. As a side note, the term "$\gamma^{\sup}$" in the convergence estimate may seem somewhat puzzling at first, since it can be negative. So, one may wonder if, contrary to intuition, large negative corrections may, in fact, help to achieve better convergence. This clearly isn't so: since $\gamma^{\sup} \ge \gamma^*$, if $\gamma^{\sup} < 0$, then $\bar{\gamma} = -\gamma^* > 0$. Furthermore, $\bar{\gamma} + \gamma^{\sup} \ge 0$ (and a fortiori $\bar{\gamma}/\alpha^* + \gamma^{\sup} \ge 0$); thus, the more negative γ_k becomes, the worse the final accuracy bound is.

Point ii. Using (4.5) in the first part of (4.14) and (4.9), one has for large enough i,

$$f(x_{k_i}) \le f^* + \sigma_{k_i} + \left(\frac{1-\alpha_{k_i}}{\alpha_{k_i}}\right)\bar{\sigma}_{k_i-1} + \frac{\varepsilon}{2} \le f^* + \sigma^*\left(1 + (1-\xi)\frac{1-\alpha^*}{\alpha^*}\right) + \varepsilon,$$

which provides the desired contradiction.

Point iii. Finally, under (3.8) ($\Rightarrow \gamma_k = \sigma_k, \bar{\sigma}_k = 0$) (4.14) is $f(x_{k_i}) \le f^* + \sigma_{k_i} + \varepsilon/2$, contradicting $f^\infty \ge f^* + \sigma^* + \varepsilon$ for i large enough; this is just the special case of Point ii for $\xi = 1$. Let us now assume that $X^* \ne \emptyset$ and select any $x^* \in X^*$. From (4.11), (2.12) (which holds due to Lemma 2.4 and (2.13)), and (4.6) we have

$$\|x_h - \bar{x}\|^2 - \|x_k - \bar{x}\|^2 \le 2\sum_{j=k}^{h-1} \nu_j(f(\bar{x}) - f^* + \bar{\sigma}_j) + \sum_{j=k}^{h-1} \nu_j^2 \|d_j\|^2.$$

Using $\bar{x} = x^* (\Rightarrow f(\bar{x}) = f^*)$, $\bar{\sigma}_k = 0$, (4.10), and (4.2), the right-hand side is bounded above by a constant; therefore, $\{x_k\}$ is bounded. From the previous results, a subsequence $\{x_{k_i}\}$ exists such that $f_{k_i} \to f^\infty$; taking subsubsequences if necessary, we have a convergent sequence to some x^∞ such that $f(x^\infty) = f^\infty$. Now, using [8, Proposition 1.3] we obtain that the whole $\{x_k\}$ converges to x^∞, and, by continuity, $\{f_k\}$ converges to f^∞. This is true, in particular, if $\sigma^* = 0 \Rightarrow f^\infty = f^*$. ∎

Thus, the results for deflection-restricted approaches (with diminishing/square summable stepsizes) closely mirror those for stepsize-restricted approaches. As far as

comparison with the literature goes, the closest result is [20, Theorem 3.6], which—without deflection—ensures $f^\infty = f^*$ (and convergence of the iterates) but requires the condition $\sum_{k=1}^\infty \nu_k\varepsilon_k < \infty$. This is not a straightforward condition to impose and clearly requires $\varepsilon_k \to 0 \Rightarrow \sigma_k \to 0$ "at least as fast" as $\nu_k \to 0$. By contrast, our Theorem 4.5 allows more relaxed conditions on both the asymptotic error and the correction. Of course, our condition (4.2) requires knowledge of f^* and is therefore, in general, not readily implementable. However, as in the stepsize-restricted case, the "basic" Theorem 4.5 provides a convenient starting point for the analysis of the implementable, target-based approaches described next, which do not require knowledge of f^* (and not even $f^* > -\infty$) or σ^k.

4.2. Target value deflection. As in the stepsize-restricted case, the nonimplementable (4.2) can be substituted with the *target value deflection* rule

$$0 \le \zeta_k = \frac{\nu_{k-1}\|d_{k-1}\|^2}{(\, f_k - f_{lev}^k\,) + \nu_{k-1}\|d_{k-1}\|^2} \le \alpha_k \le 1, \tag{4.15}$$

whereby, f^* is approximated by the target level $f_{lev}^k = f_{ref}^k - \delta_k$. As in the stepsize-restricted case, doing away with (3.2) requires defining a feasible target $\bar f$ (cf. (3.21)) to replace f^* in (4.2); then, (4.15) can be seen as a *corrected deflection* rule with

$$\lambda_k = f_k - f_{lev}^k = f_k - \bar f - \left(f_{ref}^k - \bar f - \delta_k\right),$$

where the *actual correction* $\gamma_k = f_{ref}^k - \bar f - \delta_k$ (cf. (3.22)) is unknown. It is easy to verify that Lemma 4.1, Corollary 4.2, and Theorem 4.5 hold with $\bar f$ replacing f^* and without a need for assumption (3.2). Indeed, knowing γ_k is not required for our convergence analysis, only its relationships with σ_k, in particular, in the form of (4.8), need be worked out. These clearly depend on how f_{ref} and δ_k are updated. In this case, however, there are less technical difficulties with nonvanishing quantities, and we can use very simple update rules together with a vanishing threshold.

In fact, let us assume the "obvious" reference value update $f_{ref}^k = f_{rec}^k$ and the following simplified form of (3.26):

$$\text{either} \qquad f_{ref}^\infty = f^* = -\infty \qquad \text{or} \qquad \liminf_{k\to\infty} \delta_k = 0. \tag{4.16}$$

One quite general (and simple) way of implementing it is to choose a positive vanishing nonsummable sequence $\{\Delta_k\}$, i.e.,

$$\Delta_k > 0, \qquad \liminf_{k\to\infty} \Delta_k = 0, \qquad \sum_{k=1}^\infty \Delta_k = \infty \tag{4.17}$$

and to use the threshold update rule

$$\delta_{k+1} \in \begin{cases} [\, \Delta_{r(k+1)} \,,\, \infty\,) & \text{if } f_{k+1} \le f_{lev}^k, \\ \{\Delta_{k+1}\} & \text{if } f_{k+1} > f_{lev}^k, \end{cases}$$

where $r(k)$ is the number of "resets," i.e., iterations where $f(x_{k+1}) \le f_{lev}^k$ occur, prior to iteration k. It is immediate to prove that the above implementation satisfies (4.16). In fact, let $\mathcal{R}$ be the set of resets: we have

$$f_{ref}^\infty \le f(x_1) - \sum_{k\in\mathcal{R}} \delta_k \le f(x_1) - \sum_{k\in\mathcal{R}} \Delta_{r(k)}.$$

Now, if $\mathcal{R}$ is infinite, then due to (4.17) we have $f^{\infty}_{ref} = -\infty$; otherwise, after the last iteration in $\mathcal{R}$ we have $\delta_k = \Delta_k$, and therefore, $\liminf_{k\to\infty} \delta_k = 0$. Then, the following convergence result can be proven.

THEOREM 4.6. *Under conditions* (4.10) *and* (4.1), *the algorithm employing the deflection rule* (4.15) *with threshold condition* (4.16) *attains either* $f^{\infty}_{ref} = -\infty = f^*$ *or* $f^{\infty}_{ref} \leq f^* + \sigma^*$.

Proof. If $f^{\infty} = -\infty$, there is nothing to prove, otherwise, from (3.20) and (4.16) we have $\gamma^* = f^{\infty}_{ref} - \bar{f}$ (cf. (3.27)). Assume by contradiction that $\gamma^* = f^{\infty}_{ref} - \bar{f} > \sigma^*$ $\Rightarrow f^{\infty} - \bar{f} > \sigma^*$ as $f^{\infty} \geq f^{\infty}_{ref}$. From the definitions, for all $\varepsilon > 0$ and large enough k,

$$\gamma_k \geq \gamma^* - \varepsilon \qquad \text{and} \qquad \sigma_k \leq \sigma^* + \varepsilon.$$

Hence, for $\varepsilon = (\gamma^* - \sigma^*)/2$ we have

$$\gamma^* - \varepsilon = \gamma^* - (\gamma^* - \sigma^*)/2 = (\gamma^* + \sigma^*)/2 = \sigma^* + (\gamma^* - \sigma^*)/2 = \sigma^* + \varepsilon.$$

That is, for large enough k,

$$\gamma_k \geq \gamma^* - \varepsilon = \sigma^* + \varepsilon \geq \sigma_k,$$

i.e., (4.8) holds with $\xi = 1$. Whence Theorem 4.5(ii) (with $\xi = 1$, and $\bar{f}$ replacing f^*) gives $f^{\infty} \leq \bar{f} + \sigma^*$, a contradiction. This immediately shows that $f^* = -\infty \Rightarrow$ $f^{\infty}_{ref} = -\infty$ (argue by contradiction and take $\bar{f}$ small enough). ∎

The target value deflection rules proposed in this paragraph are an implementable version of the "abstract" (4.2), which seems to be entirely new; indeed, to the best of our knowledge, there are very few comparable results in the literature. The only related result is that of the recent and independent [32], which, however, uses a different form of subgradient iteration by projecting $x_k - d_k$ onto X *before* applying the stepsize, rather than projecting d_k on the normal cone. This algorithm can attain convergence under (4.1) without using any information about the target level, but at the cost of constantly setting $\alpha_k = a\nu_k$ for a fixed constant $a > 0$, which, in particular, yields $\alpha_k \to 0$.

5. Conclusions and directions for future work.

5.1. Impact on Lagrangian optimization. We now discuss the relevance of the above analysis in a specific application, *Lagrangian relaxation* (e.g., [25, 16, 15]). There, one has a "difficult" problem

$$\sup_u \{\, c(u) \;:\; h(u) = 0\,,\; u \in U \,\} \tag{5.1}$$

which "exhibits structure," in the sense that replacing the "complicating constraints" $h(u) = 0$ with a Lagrangian term in the objective function, weighted by a vector of *Lagrangian multipliers* x, yields a Lagrangian subproblem

$$f(x) \;=\; \sup_u \{\, c(u) + \langle x, h(u)\rangle \;:\; u \in U \,\}, \tag{5.2}$$

which is "substantially easier" to solve than (5.1). Solving the corresponding *Lagrangian dual* (1.1) with this f provides an *upper bound* on the optimal value of (5.1). It is well known that the upper bound is not, in general, exact; *augmented Lagrangians* are required for obtaining a zero-gap dual. While modified subgradient methods have been proposed for these (e.g., [6]), they require the solution of a different problem

than (5.2), with a further nonlinear term in the objective function; it is often the case that these problems are considerably more difficult in practice than standard Lagrangians, especially when $h(u) = 0$ are "coupling" constraints, linking otherwise separate subsets of the variables. Assuming that only the standard Lagrangian (5.2) is solvable in practice, the (repeated, efficient) solution of (1.1) need be integrated in *enumerative approaches* [15], as follows. A feasible solution $\bar{u}$ to (5.1) is known, providing a *lower bound* $c(\bar{u})$ on the optimal value of the problem; for a given optimality tolerance ε, one is interested in determining whether or not $f^* \leq t^* = (1+\varepsilon)c(\bar{u})$. If so, $\bar{u}$ is deemed "accurate enough" a solution, and (5.1) is considered solved; otherwise, *branching* occurs where U is subdivided into a number of smallest sets, and the process is recursively repeated until the upper bound computed on each subproblem is within the prescribed tolerance from the current lower bound, which may also be improved. To avoid any unnecessary complication of the notation, we will describe the case for (5.1) and (5.2), with the only provision that, in general, one may not even have $f^* \geq c(\bar{u})$, as it is the case at the "root node."

Often, the optimization problem in (5.2) is considered "easily solvable"; this means that one assumes, given x, to be able to find an optimal solution $u_x^* \in U$ to (5.2), which gives *both* the function value $f(x) = c(u_x^*)$ *and* the subgradient $g = h(u_x^*)$. However, in order for the upper bound f^* to be "tight," it is often advisable to choose as (5.2) a problem that is *not* easy; when (5.1) is an integer linear program, for instance, choosing a U with the *integrality property* (which leads to an easy Lagrangian relaxation) provides the same bound as the ordinary continuous relaxation, which may be too weak [25, 16, 15]. Thus, it is often necessary, for the whole approach to be effective, to resort to "difficult" Lagrangian subproblems which, although easier than (5.1) in practice, fall on the same theoretical complexity class; a nice example of this can be found, e.g., in [4]. In particular, (5.2) itself may require an enumerative approach to be solved; that is, given x one may only be able to derive an upper bound $f^+(x) \geq f(x)$ (through a further relaxation of (5.2)) and some feasible but not necessarily optimal solution $u_x^- \in U$ to the problem, which produces a lower bound $c(u_x^-) = f^-(x) \leq f(x)$, together with the approximate subgradient $g = h(u_x^-)$. By enumeration, $f^+(x)$ and $f^-(x)$ can be drawn arbitrarily close together; however, this may be rather costly, especially if the *required gap* $\sigma = f^+(x) - f^-(x)$ needs to be very small (note that σ is actually the sum of two components, an upper bound error $f^+(x) - f(x)$ and a lower bound error $f(x) - f^-(x)$, but the two contributions are, in general, difficult to distinguish).

It therefore makes computational sense to consider schemes where f is only approximately computed, especially at the early stages of the algorithm for solving (1.1); this was one of the main drivers toward the development of solution approaches to (1.1) capable of dealing with approximated oracles for f [20, 21, 19]. However, not much is known about how the approximation should be chosen and possibly how the choice should evolve along with the iterates of the algorithm. The analysis in this paper, since it reveals in details the relationships between the error in the function computation, the correction in the stepsize/deflection formulae, and the resulting final accuracy of the approach, may help in deriving some first results about this issue, at least for the subgradient approaches covered by this convergence theory. Note that, for the application of interest here, the function value produced by the approximate oracle can only be taken as to be the *upper bound* $f^+(x)$ on the true value $f(x)$, which is, in general, unknown; this is because $f(x)$ is computed for *upper bounding* purposes, so only an upper bound on the true value (if small enough) can be used to certify the (approximate) optimality of a solution. The following results easily follow

from the previous analysis:

- For a subgradient scheme employing either (3.1) or (4.2), the target-level schemes (in particular, vanishing ones) predict that in order to guarantee convergence at t^* (provided that $f^* \leq t^*$ in the first place), one must have $\sigma^* \leq t^* - f^*$.
- The same holds for the (somewhat simpler) *fixed target* approach where $\lambda_k = f_k - (t^* + \delta^k) \Rightarrow \gamma_k = t^* + \delta^k - f^*$ for $\liminf_{k\to\infty} \delta_k = 0$; in fact, $f^* \leq t^*$ and $\sigma^* = t^* - f^*$ imply $f^\infty \leq t^*$. This comes directly from Theorem 3.5(ii) for (3.1) and Theorem 4.5(ii) for (4.2) (with $\xi = 1$), since $\gamma^* \geq t^* - f^* \geq \sigma^*$.
- No fixed target approach can attain convergence to t^* with lower asymptotic oracle precision. In fact, let $t_k = f_k - \lambda_k$ be the target used (in either (3.1) or (4.2)) at the kth iteration, and let $t_\infty = \liminf_{k\to\infty} t_k$. If $t_\infty > t^*$, then it is possible that for all k large enough, $f_k < t_\infty \Rightarrow \lambda_k < 0$; thus, the algorithm may "get stuck" at some iteration k, with $f_k > t^*$. If, instead, $t_\infty < t^*$, then if $t_\infty < f^*$ as well, one has $\gamma_k = t_k - f^* < 0$ for all k large enough, and therefore, $\gamma^* = t_\infty - f^* < 0$; Theorem 3.5(i) or Theorem 4.5(i) with $\sigma^* = t^* - f^*$ do *not* guarantee convergence to t^* in this case.

The above results are largely theoretical; in order to bound the "optimal" error, one would need to know the value of f^*, which is unknown. However, they indicate that the required accuracy is related to the gap between t^* and f^*, and therefore, that the worst case is when f^* is very near to the target. Thus, schemes that try to estimate f^* and revise the estimate dynamically seem to be needed for properly choosing the oracle error. More in general, the results clearly imply that *setting an asymptotic accuracy in the oracle greater than the desired final accuracy in the bound computation is, in principle, wasted*: if ε is the required final accuracy, then all the algorithms proposed herein will attain it provided that $\sigma^* = \varepsilon$. While this makes a lot of sense intuitively, we are not aware of any previous statement of this kind.

5.2. Conclusions. The contributions of the present paper are the following:

- the first convergence proofs for *approximate* subgradient algorithms combining *deflection* and *projection*, with up to *seven* different options;
- the new definition of deflection-restricted approaches;
- the new definition of *corrected* stepsize and deflection rules, and a thorough analysis of how correction impacts the asymptotic precision attained by the algorithms;
- implementable target-like versions for all algorithms.

In our opinion, one of the most interesting—although somewhat obvious, in hindsight—findings is that when dealing with inexact oracles, *an estimate of the oracle error σ_k can be useful*. Indeed, the "exact corrections" $\gamma_k = \sigma_k$ is only applicable given a "more advanced" oracle, which not only provides f_k and g_k but also σ_k; however, as discussed in the previous paragraph, such an oracle is, indeed, available in applications, e.g., related to Lagrangian relaxation. In plain words, the results in this paper suggest that if errors are made, then it is $f^* + \sigma^*$—the *lowest attainable upper bound* on f^*— that is the, and therefore, should be used as, "true target" of the approach. Using f^* instead, i.e., pretending that each g_k is a "true" subgradient rather than a σ_k-subgradient, means "aiming lower than the true target," and this hurts the convergence properties of the approach.

Clearly, there is ample scope for improvements to the obtained results. The weak boundedness condition (4.10) is satisfied by several classes of functions relevant for applications, such as Lagrangian functions of integer programs [16, 15] with compact

domains. In general, *bounding strategies* akin to those of [20, section 6] could be used to replace it; these, however, require finite upper bounds on the maximum error ε_k which, in light of our (3.6) and (4.5), do not look trivial to attain. Also, extending the present approach to *incremental methods* [20, section 9], [35, 27] would be very interesting but looks far from trivial. Deriving complexity estimates a la [20, section 8] on the different algorithms and relating them to the properties of the error sequence $\{\sigma_k\}$, the deflection sequence $\{\alpha_k\}$, and the stepsize sequences $\{\beta_k\}/\{\nu_k\}$ would clearly be interesting, as it is studying the impact of the choice of $\{\alpha_k\}$ on dual convergence properties of the approach [1, 32], e.g., borrowing from the scheme recently proposed in [29]. Finally, computational experiences are needed to assess the practical significance of the newly developed approaches. Preliminary computational results presented in [10] show that projecting the direction within a conditional subgradient scheme can be beneficial to performances; however, this is not always the case. The results obtained so far offer little guidance on the conditions under which the deflected-conditional approaches could be reliably expected to outperform the previously developed ones, as well as on other practically relevant issues, such as which of the seven projection schemes (1.5)/(1.6) is more promising in practice. An especially interesting question, from the computational viewpoint, is whether or not these methods, indeed, have similar numerical behavior as "heavy ball" ones (cf. Remark 4 after Lemma 2.6), that are quite popular in the context of training neural networks. All this calls for further study of the matter, both theoretical and computational, which we intend to pursue in the future.

Acknowledgments. We are grateful to the anonymous referees whose insightful comments have helped us to improve on the original version of the paper.

REFERENCES

[1] K. Anstreicher and L. Wolsey, *Two "well-known" properties of subgradient optimization*, Math. Program., to appear.

[2] L. Bahiense, N. Maculan, and C. Sagastizábal, *The volume algorithm revisited: Relation with bundle methods*, Math. Program., 94 (2002), pp. 41–69.

[3] F. Barahona and R. Anbil, *The volume algorithm: Producing primal solutions with a subgradient method*, Math. Program., 87 (2000), pp. 385–399.

[4] C. Beltran, C. Tadonky, and J.-P. Vial, *Solving the p-median problem with a semi-Lagrangian relaxation*, Comput. Optim. Appl., 35 (2006), pp. 239–260.

[5] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[6] R. Burachik, R. Gasimov, N. Ismayilova, and C. Kaya, *On a modified subgradient algorithm for dual problems via sharp augmented Lagrangian*, J. Global Optim., 34 (2006), pp. 55–78.

[7] P. Camerini, L. Fratta, and F. Maffioli, *On improving relaxation methods by modified gradient techniques*, Math. Program. Study, 3 (1975), pp. 26–34.

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

[9] T. Crainic, A. Frangioni, and B. Gendron, *Bundle-based relaxation methods for multicommodity capacitated fixed charge network design*, Discrete Appl. Math., 112 (2001), pp. 73–99.

[10] G. d'Antonio, *Porting ed Estensione di Codice C++ per l'Ottimizzazione non Differenziabile*, Master's Thesis, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 2006, http://www.di.unipi.it/optimize/Theses.

[11] O. du Merle, J.-L. Goffin, and J.-P. Vial, *On improvements to the analytic center cutting plane method*, Comput. Optim. Appl., 11 (1998), pp. 37–52.

[12] A. Frangioni, A. Lodi, and G. Rinaldi, *New approaches for optimizing over the semimetric polytope*, Math. Program., 104 (2005), pp. 375–388.

[13] A. Frangioni, *Solving semidefinite quadratic problems within nonsmooth optimization algorithms*, Comput. Oper. Res., 21 (1996), pp. 1099–1118.

[14] A. Frangioni, *Generalized bundle methods*, SIAM J. Optim., 13 (2002), pp. 117–156.
[15] A. Frangioni, *About Lagrangian methods in integer optimization*, Ann. Oper. Res., 139 (2005), pp. 163–193.
[16] M. Guignard, *Lagrangean relaxation*, Top, 11 (2003), pp. 151–228.
[17] B. Guta, *Subgradient Optimization Methods in Integer Programming with an Application to a Radiation Therapy Problem*, Ph.D. thesis, Teknishe Universitat Kaiserlautern, Kaiserlautern, 2003.
[18] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II—*Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.
[19] K. Kiwiel and C. Lemaréchal, *An inexact bundle variant suited to column generation*, Math. Program., to appear.
[20] K. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2004), pp. 807–840.
[21] K. Kiwiel, *A Proximal bundle method with approximate subgradient linearizations*, SIAM J. Optim., 16 (2006), pp. 1007–1023.
[22] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization - Theory and applications*, European J. Oper. Res., 88 (1996), pp. 382–403.
[23] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Ergodic, primal convergence in dual subgradient schemes for convex programming*, Math. Program., 86 (1999), pp. 283–312.
[24] 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*, European J. Oper. Res., 151 (2003), pp. 461–473.
[25] C. Lemaréchal, *Lagrangian relaxation*, in Computational Combinatorial Optimization, M. Jünger and D. Naddef, eds., Springer-Verlag, Heidelberg, 2001, pp. 115–160.
[26] C. Lim and H. Sherali, *Convergence and computational analyses for some variable target value and subgradient deflection methods*, Comput. Optim. Appl., 34 (2006), pp. 409–428.
[27] A. Nedich and D. Bertsekas, *The Effect of Deterministic Noise in Subgradient Methods*, Lab. for Information and Decision Systems Report, MIT, Cambridge, MA, 2007 (revised 2008).
[28] Y. Nesterov, *Complexity estimates of some cutting plane methods based on the analytic barrier*, Math. Program., 69 (1995), pp. 149–176.
[29] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., to appear.
[30] B. Poljak, *Introduction to Optimization*, Optimization Software, New York, 1985.
[31] B. Polyak, *Subgradient methods: A survey of Soviet research*, in Nonsmooth Optimization, Lemaréchal, C. and Mifflin, R., eds., IIASA Proceedings Series, Pergamon Press, Oxford, 1977.
[32] A. Ruszczyński, *A merit function approach to the subgradient method with averaging*, Optim. Methods Softw., 23 (2008), pp. 161–172.
[33] H. Sherali, G. Choi, and C. Tuncbilek, *A variable target value method for nondifferentiable optimization*, Oper. Res. Lett., 26 (2000), pp. 1–8.
[34] H. Sherali and C. Lim, *On embedding the volume algorithm in a variable target value method*, Oper. Res. Lett., 32 (2004), pp. 455–462.
[35] M. Solodov and S. Zavriev, *Error stability properties of generalized gradient-type algorithms*, J. Optim. Theory Appl., 98 (1998), pp. 663–680.