The Efficiency of subgradient projection methods for convex nondifferentiable optimization

Krzysztof C. Kiwiel

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The efficiency of subgradient projection methods for convex nondifferentiable optimization

Krzysztof C. KIWIEL

N° 1845
Février 1993

PROGRAMME 5

Traitement du Signal,
Automatique et
Productique

THE EFFICIENCY OF SUBGRADIENT PROJECTION METHODS FOR CONVEX NONDIFFERENTIABLE OPTIMIZATION

L'EFFICACITÉ DE LA MÉTHODE DE SOUS-GRADIENTS PAR PROJECTION, POUR L'OPTIMISATION CONVEXE NON DIFFÉRENTIABLE

Krzysztof C. KIWIEL[†]

January 19, 1993

ABSTRACT

We study subgradient methods for convex optimization that use projections onto successive approximations of level sets of the objective corresponding to estimates of the optimal value. We show that they enjoy almost optimal efficiency estimates. We present several variants, establish their efficiency estimates, and discuss possible implementations. In particular, their projection subproblems may be solved inexactly via relaxation methods, thus opening the way for parallel implementations. We discuss accelerations of relaxation methods based on simultaneous projections, surrogate constraints, and conjugate and projected (conditional) subgradient techniques.

RÉSUMÉ

Nous étudions les méthodes d'optimisation convexe qui utilisent la projection sur des approximations successives d'ensembles de niveau de la fonction-coût correspondant à des approximations de la valeur optimale. Nous montrons que ces méthodes ont une efficacité presque optimale. Nous en présentons plusieurs variantes, établissant une estimation de leur efficacité, et traitant d'implémentations possibles. En particulier, le sous-problème de projection peut être résolu de façon inexacte par des méthodes de relaxation, ce qui ouvre la possibilité d'implémentations parallèles.

[†] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland.

1 Introduction

We consider various modifications of Polyak's [Pol69] subgradient projection algorithm (SPA) and the recently proposed level method of [LNN91] for solving the convex program

$$f^* = \min\{ f(x) : x \in S \} \tag{1.1}$$

under the following assumptions. S is a nonempty compact convex subset of $\mathbb{R}^N$, f is a convex function Lipschitz continuous on S with Lipschitz constant L_f, for each $x \in S$ we can compute $f(x)$ and a subgradient $g_f(x) \in \partial f(x)$ of f at x such that $|g_f(x)| \leq L_f$, and for each $x \in \mathbb{R}^N$ we can find $P_S(x) = \arg\min\{|x - y| : y \in S\}$, its orthogonal projection on S, where $|\cdot|$ denotes the Euclidean norm.

If f^* is known, the simplest version of the SPA generates successive iterates

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f^*)g_f(x^k)/|g_f(x^k)|^2) \quad \text{for } k = 1, 2, \ldots, \tag{1.2}$$

until $g_f(x^k) = 0$, where $x^1 \in S$ and t_k are scalars in the set of *admissible stepsizes*

$$T = [\, t_{\min}, t_{\max}\,] \quad \text{for some fixed} \quad 0 < t_{\min} \leq t_{\max} < 2. \tag{1.3}$$

It has the following *efficiency estimate* for any (absolute) accuracy $\epsilon > 0$:

$$\begin{aligned} k > c_{\mathrm{SPA}}(t_{\min}, t_{\max})(\mathrm{diam}(S)L_f/\epsilon)^2 \quad &\Rightarrow \quad \min\{ f(x^j) : j = 1{:}\,k \} - f^* < \epsilon, \\ c_{\mathrm{SPA}}(t_{\min}, t_{\max}) = 1/t_{\min}(2 - t_{\max}) \quad &\text{and} \quad \min c_{\mathrm{SPA}}(\cdot,\cdot) = c_{\mathrm{SPA}}(1,1) = 1, \end{aligned} \tag{1.4}$$

where $\mathrm{diam}(S) = \sup_{x,y \in S} |x - y|$ denotes the diameter of S. This estimate (see §5) seems to be a folklore result, but it is less well known that it is optimal in a certain sense [LNN91, NeY79]: if S is a ball and $N \geq (\mathrm{diam}(S)L_f/\epsilon)^2/4$ then for any method that uses at most $(\mathrm{diam}(S)L_f/\epsilon)^2/4$ objective and subgradient evaluations there exists a function for which this method does not obtain an accuracy better than ϵ.

We present three schemes for estimating f^* in (1.2) that extend the ideas in [KAC91, KuF90, LNN91]. Two of them employ an overestimate $\bar{D} \geq \mathrm{diam}(S)$, which replaces $\mathrm{diam}(S)$ in (1.4); the third one does not involve $\bar{D}$ but is much more difficult to implement.

To enhance faster convergence, we give algorithms that use projections onto successive approximations of level sets of f derived from several accumulated subgradient linearizations of f or their *aggregates* (convex combinations) as in descent bundle methods for nondifferentiable optimization (NDO); see, e.g., [Kiw85, Lem89]. Such algorithms provide freedom to trade-off storage requirements and work per iteration for speed of convergence. Moreover, their projection subproblems can be solved efficiently even in the large-scale case by a variety of methods, especially those that can benefit from parallel computation; see, e.g., [AhC89, IDP91, Kiw92, LoH88, Oko92, Spi87, Tse90, YaM92] and the references therein. The ability to use inexact projections makes such algorithms very attractive in large-scale applications. In contrast, the existing bundle methods (see, e.g., [Kiw89, ScZ92]) employ nonstrictly convex quadratic programming (QP) subproblems, and it is not clear how to solve such QP subproblems exactly via parallel computation.

It is fruitful to view subgradient methods as extensions of relaxation methods for linear inequalities; see, e.g., [Agm54, Gof78, MoS54]. We provide a unified perspective on acceleration techniques for such methods, including simultaneous projections [Tod79], surrogate cuts [BGT81, GoT82, Oko92], dual ϵ-subgradient techniques [Brä91, KuF90, LNN91], surrogate constraints [YaM92], conjugate subgradients [CFM75, KKA87, Shc92, ShU89, Sho79], and projected (conditional) subgradients [BaG79, KiU89]. In contrast to their usual interpretations, we show that such methods hinge on *implicitly* generated affine (or polyhedral) models of f. *Explicit* use of such models allows various modifications and extensions that seem more efficient. It turns out that some of these methods are simplified versions of others that trade speed of convergence for ease of implementation. Further, our framework shows how to modify their models to account for the constraint $x \in S$. For instance, it suggests the following simple modification of (1.2)

$$x^{k+1} = \arg\min\{\, |x - x^k|^2/2 : f(x^k) + \left\langle g_f(x^k), x - x^k \right\rangle \le f^*, x \in S \,\}, \tag{1.5}$$

which seems to be more efficient in general.

In effect, we show that several versions of subgradient projection methods share efficiency estimates similar to (1.4). Since this estimate cannot, in general, be improved uniformly with respect to the dimension N by more than an absolute constant factor, all these methods are optimal in the sense of [NeY79]. We note, however, that this estimate can be attained only for really large N. We may also expect that for 'most' functions encountered in applications the methods should be much more efficient than the worst-case estimates suggest. Indeed, preliminary numerical experience with the level method of [LNN91] has been very encouraging. Yet this method is not readily implementable because it requires unbounded storage (at least of order $k(N+1)$ at iteration k). Thus the main aim of our work has been to derive methods which have comparable efficiency but are more easily implementable. In order to keep this paper reasonably short, we intend to provide numerical evidence elsewhere.

The paper is organized as follows. In §2 we introduce a general relaxation level algorithm. Its efficiency is analyzed in §3. In §4 we extend the nested ball principle of [Dre83]. Some useful modifications are given in §§5 and 6. Two alternative techniques for generating lower bounds f_{low}^k via fixed level gaps and full model minimizations are described in §7 and §8 respectively. Dual level methods are the subject of §9. In §10 we give conditions that allow efficient implementations via general relaxation and QP methods discussed in §§11 and 12, as well as 'cheap' surrogate projection methods developed in §13. Extensions of conjugate subgradient implementations are given in §14. In §15 we argue that subgradient relaxation should also include inequalities related to S. Finally, we have a concluding section.

We use the following notation. We denote by $\langle \cdot, \cdot \rangle$ and $|\cdot|$ respectively the usual inner product and norm in $\mathbb{R}^N$. $B(x,r) = \{y : |y - x| \le r\}$ denotes the ball with center x and radius $r \ge 0$. For $\epsilon \ge 0$, the ϵ-subdifferential of f at x is defined by $\partial_\epsilon f(x) = \{p \in \mathbb{R}^N : f(y) \ge f(x) + \langle p, y - x \rangle - \epsilon \quad \forall y \in \mathbb{R}^N\}$. We denote by ∂f the ordinary subdifferential $\partial_0 f$. The natural logarithm with base e is denoted by $\ln(\cdot)$. We let $1{:}k$ denote $1, 2, \ldots, k$. For brevity, we let $a/bc = a/(bc)$. The convex hull is denoted by co.

2 The relaxation level algorithm

In this section we describe our first modification of the SPA. As in [BaS81, KAC91, LNN91], when the optimal value f^* is unknown, it may be replaced in (1.2) by a variable target (level) value

$$f^k_{\rm lev} = f^k_{\rm up} - \kappa(f^k_{\rm up} - f^k_{\rm low}) = \kappa f^k_{\rm low} + (1-\kappa) f^k_{\rm up}, \tag{2.1}$$

where $0 < \kappa < 1$ is fixed, $f^k_{\rm up} = \min_{j=1:k} f(x^j)$ is an upper bound on f^*, and the lower bound $f^k_{\rm low} \le f^*$ is chosen to ensure $f^k_{\rm lev} \to f^*$ as $k \to \infty$. Thus we obtain the subgradient projection level algorithm (SPLA):

$$x^{k+1} = P_S(x^k - t_k(f(x^k) - f^k_{\rm lev}) g_f(x^k)/|g_f(x^k)|^2) \quad \text{for } k = 1,2,\ldots; \tag{2.2}$$

if $g_f(x^k) = 0$ then, of course, the method stops with an optimal x^k in $S^* = \operatorname{Arg\,min}_S f$. To get some feeling about possible updates of $f^k_{\rm low}$, it is instructive to consider first the following ideal bisection method (cf. [MTA81]).

Algorithm 2.1 (*ideal level method for* (1.1)).
Step 0. Choose $0 < \kappa < 1$, $z^1 \in S$ and $f^1_{\rm low} \le f^*$. Set $k = 1$.

Step 1. Set $f^k_{\rm up} = f(z^k)$, $f^k_{\rm lev}$ by (2.1) and the optimality gap $\Delta^k = f^k_{\rm up} - f^k_{\rm low}$.

Step 2. Let $\mathcal{L}(f, f^k_{\rm lev}) = \{x : f(x) \le f^k_{\rm lev}\}$. If $\mathcal{L}(f, f^k_{\rm lev}) \cap S = \emptyset$, go to Step 4.

Step 3. Find $z^{k+1} \in \mathcal{L}(f, f^k_{\rm lev}) \cap S$, set $f^{k+1}_{\rm low} = f^k_{\rm low}$, increase k by 1 and go to Step 1.

Step 4. Set $z^{k+1} = z^k$, $f^{k+1}_{\rm low} = f^k_{\rm lev}$, increase k by 1 and go to Step 1.

Clearly, the method produces $f^k_{\rm low} \le f^*$, $f^k_{\rm up} - f^* \le \Delta^k$ and $\Delta^{k+1} \le \max\{\kappa, (1-\kappa)\}\Delta^k$ for all k. The crucial property is that $\mathcal{L}(f, f^k_{\rm lev}) \cap S = \emptyset$ implies $f^k_{\rm lev} < f^*$.

To make Algorithm 2.1 implementable, we need a submethod for finding a point in $\mathcal{L}(f, f^k_{\rm lev}) \cap S$ or detecting that $\mathcal{L}(f, f^k_{\rm lev}) \cap S = \emptyset$. For this kth *set intersection problem*, an iteration of the successive projections method [GPR67] of the form $\hat{x}^{k+1} = P_S(P_{\mathcal{L}(f, f^k_{\rm lev})}(x^k))$ can be implemented approximately as follows. Letting $\bar{f}(\cdot; y) = f(y) + \langle g_f(y), \cdot - y\rangle$ denote the *linearization* of f at any $y \in S$, with $\bar{f}(\cdot; y) \le f$ and $\bar{f}(y; y) = f(y)$ by convexity, we have

$$S^* = \{\, x : \bar{f}(x; y) \le f^* \ \forall y \in S \,\} \cap S \tag{2.3}$$

and

$$\mathcal{L}(f, f^k_{\rm lev}) = \{\, x : \bar{f}(x; y) \le f^k_{\rm lev} \ \forall y \in S \,\}. \tag{2.4}$$

We may use some accumulated linearizations $f^j = \bar{f}(\cdot; x^j)$, $j \le k$, in the kth model of f

$$\hat{f}^k(x) = \max\{\, f^j(x) : j \in J^k \,\} \quad \text{with} \quad k \in J^k \subset \{1{:}\,k\} \tag{2.5}$$

and let

$$x^{k+1} = P_S(x^k + t_k[P_{\mathcal{L}(\hat{f}^k, f^k_{\rm lev})}(x^k) - x^k]), \tag{2.6}$$

where we have *underprojection* if $t_k < 1$ or *overprojection* if $t_k > 1$. For instance, (2.6) gives (1.2) when $f^k_{\rm lev} = f^*$, $J^k = \{k\}$ and $\mathcal{L}(\hat{f}^k, f^k_{\rm lev})$ is the halfspace $H^k = \{x : f^k(x) \le f^k_{\rm lev}\}$

given by the inequality of (2.3) most violated at x^k; of course, $P_{H^k}(x^k) = x^k - (f(x^k) - f^k_{\rm lev})g_f(x^k)/|g_f(x^k)|^2$. This is just an iteration of a relaxation method for solving the inequalities of (2.4), followed by a projection on S. As for (2.2), $f^k_{\rm low}$ *may be increased to* $f^k_{\rm lev}$ *when it is discovered that these inequalities do not have a solution in* S. Hence we shall exploit the fact that certain versions of successive projections methods can detect in finite time that a given set intersection problem is unsolvable (although they need not find a solution in finite time when it exists). As will be seen below, the main idea of such methods is to reduce the distance of the iterates from S^*. They may be painfully slow, even in the most favorable case of $f^k_{\rm lev} \doteq f^*$, when only one inequality of (2.3) is considered at a time. To accelerate convergence, we may use a larger J^k, i.e., a tighter approximation $\hat{f}^k$ of f.

To illustrate these facts we need a result of Agmon [Agm54]. Given a closed convex set $C \subset \mathbb{R}^N$ and an admissible stepsize $t \in T$, we define the *relaxation operator*

$$\mathcal{R}_{C,t}(x) = x + t(P_C(x) - x) \tag{2.7}$$

(where $P_C(x) = x$ if $C = \emptyset$) that has the Fejér contraction property

$$\begin{aligned} |y - \mathcal{R}_{C,t}(x)|^2 &\leq |y - x|^2 - t(2-t)|x - P_C(x)|^2 \\ &\leq |y - x|^2 - t_{\min}(2 - t_{\max})d_C^2(x) \quad \forall y \in C, x \in \mathbb{R}^N, \end{aligned} \tag{2.8}$$

where $d_C(x) = \inf_{y \in C} |x - y|$. Indeed, if $y \in C$, $P = P_C$ and $z = x + t(P(x) - x)$ then

$$\begin{aligned} |y - z|^2 &= |y - x|^2 + (t|P(x) - x|)^2 - 2t\,\langle y - x, P(x) - x\rangle \\ &= |y - x|^2 + (t|P(x) - x|)^2 - 2t\,\langle P(x) - x, P(x) - x\rangle - 2t\,\langle y - P(x), P(x) - x\rangle \\ &\leq |y - x|^2 - t(2-t)|P(x) - x|^2 \leq |y - x|^2 - t_{\min}(2 - t_{\max})|P(x) - x|^2 \end{aligned}$$

from the projection property $\langle y - P(x), P(x) - x\rangle \geq 0$ and (1.3). Note that $t_{\min}(2 - t_{\max})$ in (2.8) can be replaced by $\min_{t \in T} t(2-t)$.

Figure 2.1 illustrates the Fejér property of (2.2) with $H^k = \mathcal{L}(f^k, f^k_{\rm lev})$. For motivation, we now state some facts that will be proved later. Suppose we have generated some $r_k \geq d_{S^*}(x^k)$ (starting, e.g., from $r_1 = \bar{D} \geq \mathrm{diam}(S)$), so that $B(x^k, r_k) \cap S^* \neq \emptyset$. If $f^k_{\rm lev} \geq f^*$ then $S^* \subset H^k$, so setting $y^k = P_{H^k}(x^k)$, finding r_{k+1} from

$$r_{k+1}^2 = r_k^2 - t_k(2 - t_k)|y^k - x^k|^2 \tag{2.9}$$

and applying (2.8) twice we deduce that $S^* \cap B(x^k, r_k) \subset S^* \cap B(x^{k+1}, r_{k+1})$. Thus we improve our localization of the solution (since $r_{k+1} < r_k$ due to $x^k \notin H^k$ from $f(x^k) > f^k_{\rm lev}$). On the other hand, if $t_k(2 - t_k)d^2_{H^k}(x^k) > r_k^2$ then $f^k_{\rm lev} < f^*$ (by contradiction), so we may increase $f^k_{\rm low}$ to $f^k_{\rm lev}$ and reset r_{k+1} to $\bar{D}$. To sum up, if $f^k_{\rm lev} \geq f^*$ then progress towards the solution is measured by the magnitude of $d_{H^k}(x^k)$, otherwise $d_{H^k}(x^k)$ may be used to shrink r_k until $f^k_{\rm lev} < f^*$ is discovered; thus $d_{H^k}(x^k)$ should be as large as possible in both cases. Hence the algorithm may be accelerated by choosing a smaller $\mathcal{L}(\hat{f}^k, f^k_{\rm lev})$ to produce $d_{\mathcal{L}(\hat{f}^k, f^k_{\rm lev})}(x^k) > d_{H^k}(x^k)$. However, a large J^k in (2.5) would create difficulties with storage and work per iteration. This raises the following basic questions. Is it possible to select J^k so that $\hat{f}^k$ approximates f tightly in the region of interest without J^k becoming inordinately

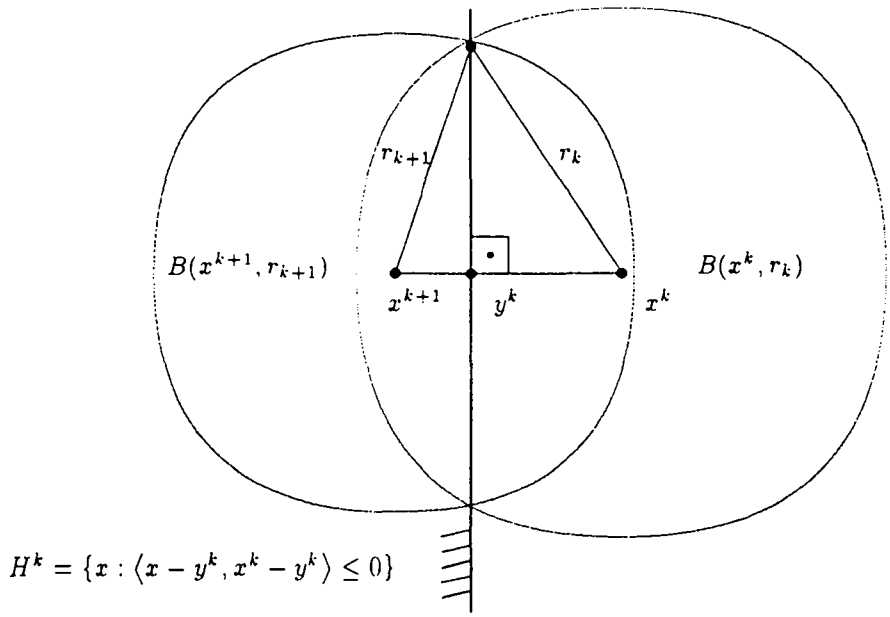

Figure 2.1: Illustration of the Fejér property of $x^{k+1} = x^k + t_k(y^k - x^k)$ with $y^k = P_{H^k}(x^k)$ when $x^{k+1} \in S$. By Pythagoras' theorem, $r_{k+1}^2 - |x^{k+1} - y^k|^2 = r_k^2 - |y^k - x^k|^2$, where $x^{k+1} - y^k = (t_k - 1)(y^k - x^k)$, so $r_{k+1}^2 = r_k^2 - t_k(2 - t_k)|y^k - x^k|^2$. Clearly, $B(x^k, r_k) \cap H^k \subset B(x^{k+1}, r_{k+1}) \cap H^k$.

large? Can we reduce J^k by replacing some f^j with their convex combinations, i.e., by aggregating some constraints in $\mathcal{L}(\hat{f}^k, f_{\text{lev}}^k)$? Should not $\mathcal{L}(\hat{f}^k, f_{\text{lev}}^k)$ be augmented with some inequalities related to S? Instead of finding $P_{\mathcal{L}(\hat{f}^k, f_{\text{lev}}^k)}(x^k)$, can we perform several 'simpler' projections (possibly inexactly and in parallel) and combine their solutions? Our partial answers to these questions will involve a combination of some quite technical properties of relaxation methods. For instance, note that, in view of the outer projection in (2.2), r_{k+1}^2 in (2.9) could be further reduced by $d_S^2(z^k)$, where $z^k = x^k + t_k(y^k - x^k)$. In fact more than two successive projections could be employed to reduce r_{k+1}. We shall need rather abstract notation to make such concepts precise.

Let δ_S denote the indicator of S ($\delta_S(x) = 0$ if $x \in S$, ∞ otherwise) and $f_S = f + \delta_S$ the extended objective. Let $\check{f}^k = \max_{j=1:k} f^j$ denote the kth 'best' model of f (which we would *not* like to store). Note that $\check{f}_S^k = \check{f}^k + \delta_S$ is the largest convex minorant of f_S compatible with the accumulated information about f. Clearly, f^k, $\hat{f}^k$, $\check{f}^k$ and $\check{f}_S^k$ belong to the following *set of admissible models of* f_S

$$\Phi = \{\, \phi \colon \mathbb{R}^N \to (-\infty, \infty] : \phi \text{ is closed convex and } \phi(x) \leq f^* \ \forall x \in S^* \,\}. \tag{2.10}$$

At iteration k, we may choose a model $\phi^k \in \Phi$ such that $\phi^k \geq f^k$ (to exploit the latest subgradient information), and a stepsize $t_k \in T$. Then the iteration

$$x^{k+1} = P_S(\mathcal{R}_{\mathcal{L}(\phi^k, f_{\text{lev}}^k), t_k}(x^k)) \tag{2.11}$$

is a generalization of (2.2) and (2.6), which have $\phi^k = f^k$ and $\phi^k = \hat{f}^k$ respectively. This notation is convenient for the implementations discussed later, in which each ϕ^k may be the maximum of several accumulated linearizations f^j, $j \leq k$, or their convex combinations, possibly augmented with δ_S or its convex minorants. It will also prepare ground for extensions which use several models from Φ at each iteration for successive or parallel relaxations. (For the first reading, one may assume $\phi^k = f^k$.) We should, of course, ensure that $\mathcal{L}(\phi^k, f^k_{\text{lev}}) \neq \emptyset$ in (2.11) (detecting this may require calculating $\inf \phi^k$ approximately). Since, by (2.10),

$$\emptyset \neq S^* \subset \mathcal{L}(\phi, f^k_{\text{lev}}) \quad \text{if} \quad \phi \in \Phi \text{ and } f^k_{\text{lev}} \geq f^*, \tag{2.12}$$

$\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$ means we may repeat (2.11) with f^k_{low} increased to f^k_{lev}. Note that this cannot happen in the simplest method (2.2), where the test based on r_k must be employed.

We may now state the first general subgradient projection algorithm with relaxation and target level updating. Its notation is slightly redundant, being geared towards subsequent convergence proofs and modifications.

Algorithm 2.2.

Step 0 (*Initialization*). Select an initial point $x^1 \in S$, a final optimality tolerance $\epsilon_{\text{opt}} \geq 0$, a level parameter $0 < \kappa < 1$, and stepsize parameters $0 < t_{\min} \leq t_{\max} < 2$. Choose $\bar{D} \geq \text{diam}(S)$ and $f^1_{\text{low}} \leq f^*$. Set $\rho_1 = 0$ and $f^0_{\text{up}} = \infty$. Set the counters $k = 1$, $l = 0$ and $k(0) = 0$ ($k(l)$ will denote the iteration number of the lth increase of f^k_{low}).

Step 1 (*Objective evaluation*). Calculate $f(x^k)$ and $g_f(x^k)$.

Step 2 (*Level update*). Set $f^k_{\text{up}} = \min\{f(x^k), f^{k-1}_{\text{up}}\}$, f^k_{lev} by (2.1) and the *gap* $\Delta^k = f^k_{\text{up}} - f^k_{\text{low}}$.

Step 3 (*Stopping criterion*). If $\min\{\Delta^k, |g_f(x^k)|/\bar{D}\} \leq \epsilon_{\text{opt}}$, terminate.

Step 4 (*Projections*). Perform (2.11), checking if it is well-defined, as follows:

(i) Choose an admissible model $\phi^k \in \Phi$ such that $\phi^k \geq f^k$ and a stepsize $t_k \in T$.

(ii) If $\mathcal{L}(\phi^k, f^k_{\text{lev}}) = \emptyset$ go to Step 5. Otherwise, set $y^k = P_{\mathcal{L}(\phi^k, f^k_{\text{lev}})}(x^k)$, $z^k = x^k + t_k(y^k - x^k)$, $x^{k+1} = P_S(z^k)$, $\rho^k_\phi = t_k(2 - t_k)|y^k - x^k|^2$ and $\rho^k_S = |x^{k+1} - z^k|^2$.

(iii) If $\rho_k + \rho^k_\phi + \rho^k_S > \bar{D}^2$, go to Step 5; otherwise, go to Step 6.

Step 5 (*Update lower bound*).

(i) Choose a lower bound $\hat{f}^k_{\text{low}} \in [\max\{f^k_{\text{low}}, f^k_{\text{lev}}\}, f^*]$ (e.g., $\hat{f}^k_{\text{low}} = \max\{f^k_{\text{low}}, f^k_{\text{lev}}\}$). Set $f^{k+1}_{\text{low}} = \hat{f}^k_{\text{low}}$, $\rho_{k+1} = 0$ and $\hat{\Delta}^k = f^k_{\text{up}} - \hat{f}^k_{\text{low}}$.

(ii) If $\hat{\Delta}^k \leq \epsilon_{\text{opt}}$, terminate; otherwise, continue.

(iii) Set $x^{k+1} = x^k$ (*null step*), $k(l+1) = k$, and increase k and l by 1. Go to Step 2.

Step 6 (*Serious step*). Set $f^{k+1}_{\text{low}} = \hat{f}^k_{\text{low}} = f^k_{\text{low}}$, $\hat{\Delta}^k = \Delta^k$ and $\rho_{k+1} = \rho_k + \rho^k_\phi + \rho^k_S$. Increase k by 1 and go to Step 1.

A few comments on the method are in order.

At Step 0, f^1_{low} may be obtained, e.g., from a relaxation of (1.1), or from the relations

$$f^* \geq \min_S f^1 \geq f(x^1) - |g_f(x^1)| \,\text{diam}(S) \geq f(x^1) - |g_f(x^1)|\bar{D}, \tag{2.13}$$

since $f \geq f^1 = f(x^1) + \langle g_f(x^1), \cdot - x^1 \rangle \geq f(x^1) - |g_f(x^1)||\cdot - x^1|$ by the Cauchy-Schwarz inequality. In many applications one may find a 'simple' set (e.g., a box or a ball) that

contains S; the diameter of this set may serve as $\bar{D}$. (Choosing $f^1_{\rm low}$ and $\bar{D}$ when f is strongly convex on S is discussed in [KAC91]; see also [KuF90].) In general, the algorithm should perform better the closer $f^1_{\rm low}$ and $\bar{D}$ are to f^* and $\mathrm{diam}(S)$, respectively.

Note that the f-evaluation at Step 1 is skipped if $x^k = x^{k-1}$ after a null step at Step 5, i.e., if $k = k(l) + 1$. The current number of f-evaluations is $k - l$.

Step 3 is justified by the optimality estimates (2.14) and $f^* \geq f(x^k) - |g_f(x^k)|\,\mathrm{diam}(S)$.

Step 4 performs the two successive relaxations of (2.11), unless an exit to Step 5 occurs with $f^k_{\rm lev} < f^*$. The exit from Step 4(ii) is justified by (2.12), and from Step 4(iii) by Lemma 3.3, which formalizes our argument concerning (2.9). Specifically, with $r_k^2 = \bar{D}^2 - \rho_k$, Step 6 replaces (2.9) by $r^2_{k+1} = r_k^2 - \rho^k_\phi - \rho^k_S$, whereas Steps 0 and 5 ensure $r_{k(l)+1} = \bar{D}$.

Let us split the iterations into groups $K_0 = \{1\!:\!k(1)-1\}$ and $K_l = \{k(l)\!:\!k(l+1)-1\}$ if $l \geq 1$. Each group K_l ends by discovering that the target level is unattainable. Then an increase of the lower bound reduces the gap between the bounds by at least a fraction of $\kappa < 1$. The remaining level and gap decreases within each group occur only when the objective improves, with the lower bound staying fixed. These simple properties of the method may by derived inductively from the following observations. By construction, $f^k_{\rm up} \geq f^{k+1}_{\rm up} \geq f^* \geq f^{k+1}_{\rm low} = \hat{f}^k_{\rm low} \geq f^k_{\rm low}$, $\hat{\Delta}^k = f^k_{\rm up} - \hat{f}^k_{\rm low}$ and $\Delta^k = f^k_{\rm up} - f^k_{\rm low}$, so the gaps $\hat{\Delta}^k \leq \Delta^k$ overestimate the optimality gap:

$$f^k_{\rm up} - f^* = \min\{\, f(x^j) : j = 1\!:\!k \,\} - f^* \leq \hat{\Delta}^k \leq \Delta^k \tag{2.14}$$

and $\Delta^{k+1} \leq \hat{\Delta}^k \leq \Delta^k$ for all k. In fact, if $k(l) < k < k(l+1)$ then $f^k_{\rm low} = \hat{f}^{k(l)}_{\rm low}$ ($= f^1_{\rm low}$ if $l = 0$); therefore, the level $f^k_{\rm lev} = f^k_{\rm low} + (1-\kappa)\Delta^k$ cannot increase:

$$f^{k(l)+1}_{\rm lev} \geq f^j_{\rm lev} \geq f^k_{\rm lev} \quad \text{if} \quad k(l) < j \leq k \leq k(l+1) \tag{2.15}$$

and $\Delta^k = \hat{\Delta}^k$ if $k(l) < k < k(l+1)$. Hence $\hat{f}^k_{\rm low}$ and $\hat{\Delta}^k$ only reflect the improvement in $f^k_{\rm low}$ and Δ^k at iterations $k = k(l+1)$, $l \geq 0$. Then at Step 5, $\hat{f}^k_{\rm low} \geq f^k_{\rm lev} = f^k_{\rm up} - \kappa\Delta^k$ implies $\hat{\Delta}^k = f^k_{\rm up} - \hat{f}^k_{\rm low} \leq \kappa\Delta^k$. Thus we have the useful relations

$$\Delta^k \geq \hat{\Delta}^k \geq \hat{\Delta}^{k(l+1)}/\kappa \quad \text{if} \quad k \in K_l \text{ and } l \geq 0, \tag{2.16}$$

$$\hat{\Delta}^{k(l)} \leq \kappa^l \Delta^1 \quad \text{if} \quad l \geq 1. \tag{2.17}$$

3 Efficiency

Our aim is to show that the SPLA of (2.2) has the following efficiency estimate for any $\epsilon > 0$:

$$k > c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon)^2 \quad \Rightarrow \quad \min\{\, f(x^j) : j = 1\!:\!k \,\} - f^* < \epsilon, \tag{3.1a}$$

$$c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2(1-\kappa^2), \tag{3.1b}$$

$$\min c_{\rm SPLA}(\cdot,\cdot,\cdot) = c_{\rm SPLA}(1, 1, 1/\sqrt{2}) = 4, \tag{3.1c}$$

and to establish a modified form of this estimate for Algorithm 2.2. We assume, with no loss of generality, that the tolerance $\epsilon_{\rm opt} = 0$ and that the algorithm does not terminate, i.e., $\Delta^k \geq \hat{\Delta}^k > 0$ for all k.

We start by showing that each first relaxation at Step 4 provides a significant growth of ρ_k related to Fejér contractions. Note that with $H^k = \{x : f^k(x) \le f^k_{\rm lev}\}$ we have

$$d_C(x^k) \ge d_{H^k}(x^k) = (f(x^k) - f^k_{\rm lev})/|g_f(x^k)| \quad \text{if} \quad C \subset H^k. \tag{3.2}$$

Lemma 3.1. *If $\mathcal{L}(\phi^k, f^k_{\rm lev}) \neq \emptyset$ at Step 4 then $\rho^k_\phi \ge t_{\min}(2 - t_{\max})(\kappa\Delta^k/L_f)^2$.*

Proof. Use (3.2) with $\mathcal{L}(\phi^k, f^k_{\rm lev}) \subset H^k$ (from $\phi^k \ge f^k$), $|g_f(x^k)| \le L_f$ and $f(x^k) - f^k_{\rm lev} \ge f^k_{\rm up} - (f^k_{\rm up} - \kappa\Delta^k) = \kappa\Delta^k$ (from $f(x^k) \ge f^k_{\rm up}$), recall Step 4 and (1.3). □

Lemma 3.2. *Suppose $y \in \mathcal{L}(\phi^k, f^k_{\rm lev}) \cap S$ for iterations $k = k_1{:}\,k_2$ that do not execute Step 5 (i.e., y is a common point of all the sets involved in the successive relaxations (2.11) at Step 4 for such k). Then*

$$\rho_{k_2+1} - \rho_{k_1} = \sum_{k=k_1}^{k_2} [t_k(2 - t_k)|y^k - x^k|^2 + |x^{k+1} - z^k|^2] \le |y - x^{k_1}|^2 - |y - x^{k_2+1}|^2. \tag{3.3}$$

Proof. Fix $k \in [k_1, k_2]$. Use (2.8) with $C = \mathcal{L}(\phi^k, f^k_{\rm lev})$, $t = t_k$ and $x = x^k$, and next with $C = S$, $t = 1$ and $x = z^k$ to get

$$\rho^k_\phi = t_k(2 - t_k)|y^k - x^k|^2 \le |y - x^k|^2 - |y - z^k|^2, \tag{3.4a}$$

$$\rho^k_S = |x^{k+1} - z^k|^2 \le |y - z^k|^2 - |y - x^{k+1}|^2. \tag{3.4b}$$

Add the inequalities above to get $\rho_{k+1} - \rho_k = \rho^k_\phi + \rho^k_S \le |y - x^k|^2 - |y - x^{k+1}|^2$. Adding these inequalities for $k = k_1{:}\,k_2$ yields (3.3). □

The next result validates the test for increasing $f^k_{\rm low}$ from Step 4(iii).

Lemma 3.3. *If $f^k_{\rm lev} \ge f^*$ at Step 4 then Step 5 is not entered and*

$$\rho_{k+1} = \rho_k + \rho^k_\phi + \rho^k_S \le |y - x^{k(l)+1}|^2 - |y - x^{k+1}|^2 \le \operatorname{diam}(S)^2 \le \bar{D}^2 \quad \forall y \in S^*. \tag{3.5}$$

Proof. Since $f^k_{\rm lev} \ge f^*$, (2.12) and (2.15) imply that the assumptions of Lemma 3.2 hold for any fixed $y \in S^*$, $k_1 = k(l) + 1$ and $k_2 = k - 1$. Then, due to the rules of Steps 5 and 6, (3.3) becomes $\rho_k \le |y - x^{k(l)+1}|^2 - |y - x^k|^2$. Adding this inequality to (3.4) we get (3.5), noting that $\rho_k + \rho^k_\phi + \rho^k_S \le |y - x^{k(l)+1}|^2 \le \bar{D}^2$, i.e., no null step occurs. □

We may now estimate the rate of decrease of the gap Δ^k within each group K_l.

Lemma 3.4. *If $k(l) < k < k(l+1)$ and $\Delta^k > 0$ then*

$$k - k(l) \le (\bar{D}L_f/\kappa\Delta^k)^2/t_{\min}(2 - t_{\max}). \tag{3.6}$$

Proof. Note that $\Delta^j \ge \Delta^k$ for $j = 1{:}\,k$ because Δ^j never increases. By the rules of Steps 4 and 5, we have $\rho_{k+1} \le \bar{D}^2$ (otherwise $k(l+1) = k$ would occur, a contradiction) and

$$\begin{aligned} \bar{D}^2 \;&\ge\; \rho_{k+1} \ge \sum_{j=k(l)+1}^{k} \rho^j_\phi \ge t_{\min}(2 - t_{\max}) \sum_{j=k(l)+1}^{k} (\kappa\Delta^j/L_f)^2 \\ &\ge\; t_{\min}(2 - t_{\max})(\kappa\Delta^k/L_f)^2(k - k(l)) \end{aligned}$$

from Lemma 3.1. Rearranging, we get (3.6). □

At Step 2, let $n_f^k = k - l$ and $l^k = l$ denote the total numbers of f-evaluations and lower bound increases, respectively. In (3.7) below we in fact relate n_f^k to the gap $\hat{\Delta}^k$.

Lemma 3.5. *If $\hat{\Delta}^{k_\epsilon} \geq \epsilon > 0$ for some $k_\epsilon \in K_m$ and $m \geq 0$, then $n_f^{k_\epsilon} = k_\epsilon - m$ and*

$$k_\epsilon \leq m + (\bar{D}L_f/\epsilon)^2/\kappa^2(1-\kappa^2)t_{\min}(2-t_{\max}). \tag{3.7}$$

If additionally $\Delta^1 \leq \bar{D}L_f$, then $m \leq -\ln(\bar{D}L_f/\epsilon)/\ln(\kappa)$ and

$$k_\epsilon \leq (\bar{D}L_f/\epsilon)^2[1/\kappa^2(1-\kappa^2)t_{\min}(2-t_{\max}) - 1/2e\ln(\kappa)]. \tag{3.8}$$

Proof. (i) Let $K(\epsilon) = \{1{:}\,k_\epsilon\}$. Since $\hat{\Delta}^{k_\epsilon} \geq \epsilon > 0$ and $\Delta^{k+1} \leq \hat{\Delta}^k \leq \Delta^k$ for all k, use (2.16) and induction to obtain $\Delta^k \geq \epsilon/\kappa^{m-l}$ for all $k \in K_l \cap K(\epsilon)$ and $l = 0{:}\,m$.

(ii) Let $c = (\bar{D}L_f/\kappa)^2/t_{\min}(2-t_{\max})$. By (i) and Lemma 3.4, $|K_l \cap K(\epsilon)| \leq 1 + c\kappa^{2(m-l)}/\epsilon^2$ for $l = 1{:}\,m$ and $|K_0 \cap K(\epsilon)| \leq c\kappa^{2m}/\epsilon^2$. Since $0 < \kappa < 1$, we get (3.7) from

$$k_\epsilon = \sum_{l=0}^{m} |K_l \cap K(\epsilon)| \leq m + \sum_{l=0}^{m} (c/\epsilon^2)\kappa^{2(m-l)} \leq m + c/\epsilon^2(1-\kappa^2).$$

(iii) If $\Delta^1 \leq \bar{D}L_f$ and $m > 0$, then (2.17) yields $\epsilon \leq \hat{\Delta}^{k(m)} \leq \kappa^m \bar{D}L_f$, so $m \leq -\ln(\bar{D}L_f/\epsilon)/\ln(\kappa)$ in (3.7). Thus, to get (3.8), it suffices to prove that $-\ln(t)/\ln(\kappa) \leq -t^2/2e\ln(\kappa)$ for all $t > 0$. Indeed, $t^2 - 2e\ln(t) \geq 0$ for all $t > 0$ (minimize it!). □

We may now state our principal result. Notice that, in view of (2.13), we may always ensure that $\Delta^1 \leq \bar{D}L_f$ by taking $f_{\text{low}}^1 \geq f(x^1) - |g_f(x^1)|\bar{D}$, and recall (2.14).

Theorem 3.6. *If $\Delta^1 \leq \bar{D}L_f$ then the following efficiency estimate holds for each $\epsilon > 0$:*

$$k > c_{\text{RLA}}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon)^2 \quad \Rightarrow \quad f_{\text{up}}^k - f^* \leq \hat{\Delta}^k < \epsilon, \tag{3.9a}$$

$$c_{\text{RLA}}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2-t_{\max})\kappa^2(1-\kappa^2) - 1/2e\ln(\kappa), \tag{3.9b}$$

$$\min c_{\text{RLA}}(\cdot,\cdot,\cdot) = c_{\text{RLA}}(1, 1, 0.677653\ldots) \approx 4.49950. \tag{3.9c}$$

Proof. This is an immediate consequence of Lemma 3.5. □

Let $x_{\text{rec}}^k \in \{x^j\}_{j=1}^k$ be such that $f(x_{\text{rec}}^k) = f_{\text{up}}^k$ $(= \min_{j=1:k} f(x^j))$, for all k.

Corollary 3.7. *If $\epsilon_{\text{opt}} = \epsilon > 0$ and $\Delta^1 \leq \bar{D}L_f$ then the algoritm will terminate with $f(x_{\text{rec}}^k) \leq f^* + \epsilon$ in $k = 1 + k_\epsilon$ iterations after $n_f^k = 1 + n_f^{k_\epsilon}$ f-evaluations, where k_ϵ and $n_f^{k_\epsilon} = k_\epsilon - m$ satisfy the bounds of Lemma 3.5.* □

For completeness, we include an asymptotic result.

Theorem 3.8. *If the algorithm does not terminate then f_{up}^k, f_{low}^k and f_{lev}^k converge to f^*, and Δ^k and $\hat{\Delta}^k$ converge to zero as $k \to \infty$. Moreover, $\{x_{\text{rec}}^k\}$ converges to S^*.*

Proof. Since $\hat{\Delta}^k > 0$ never increases, $\hat{\Delta}^k \downarrow 0$ either by (2.17) if $l \to \infty$ or by Lemma 3.5 otherwise (then m would be bounded in (3.7)). Hence the facts that $\Delta^{k+1} \le \hat{\Delta}^k \le \Delta^k$ and $\max\{|f^k_{\rm low} - f^*|, |f^k_{\rm up} - f^*|, |f^k_{\rm lev} - f^*|\} \le \Delta^k$ for all k imply the first assertion. The second one follows from $f(x^k_{\rm rec}) \to f^*$, the continuity of f and the compactness of S. □

Remark 3.9. In view of the preceding results, we again emphasize the crucial role of $\rho^k_\phi = t_k(2-t_k)d^2_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$ in our efficiency analysis. The algorithm may be accelerated (locally) by choosing ϕ^k and t_k to enhance a large ρ^k_ϕ. (In fact we should try to increase the less easily manageable quantity $\rho^k_\phi + \rho^k_S$ instead of just ρ^k_ϕ.) Our efficiency estimates are best when $t_{\rm min} = t_{\rm max} = 1$; also $t_k = 1$ maximize each ρ^k_ϕ. However, as in other relaxation methods, other choices of t_k may be preferable in practice.

4 The nested ball principle

We shall need the following reformulation of Lemma 3.3 in terms of $r^2_k = \bar{D}^2 - \rho_k$. It generalizes similar results of [Gof81, Tel82] obtained for classical relaxation methods.

Lemma 4.1 (The ball induction principle). *If $f^k_{\rm lev} \ge f^*$ at Step 4 then $\emptyset \ne S^* \cap B(x^k, r_k) \subset S^* \cap B(z^k, (r^2_k - \rho^k_\phi)^{1/2}) \subset S^* \cap B(x^{k+1}, r_{k+1})$.* □

The following result extends one of Drezner [Dre83] (and simplifies its proof).

Lemma 4.2 (The nested ball principle). *If $(\bar{D} - |z^k - x^{k(l)+1}|)^2 > r^2_k - \rho^k_\phi$ or $(\bar{D} - |x^{k+1} - x^{k(l)+1}|)^2 > r^2_k - \rho^k_\phi - \rho^k_S$ at Step 4 then $f^k_{\rm lev} < f^*$.*

Proof. For contradiction suppose $f^k_{\rm lev} \ge f^*$. Let $x = x^{k(l)+1}$, $z = z^k$, $\hat{r} = (r^2_k - \rho^k_\phi)^{1/2}$ and $y \in S^* \cap B(x, \bar{D}) \cap B(z, \hat{r})$ (cf. Lemma 4.1). Suppose $\bar{D} > \hat{r} + |z - x|$. By construction and (3.3)–(3.5), $|y - z|^2 \le |y - x|^2 + \hat{r}^2 - \bar{D}^2 \le (|y - z| + |z - x|)^2 + \hat{r}^2 - \bar{D}^2$ with $|z - x| \ne 0$ due to $\hat{r} < \bar{D}$, so $|y - z| \ge (\bar{D}^2 - \hat{r}^2 - |z - x|^2)/2|z - x| > \hat{r}$ contradicts $y \in B(z, \hat{r})$. Hence $\bar{D} \le \hat{r} + |z - x|$ and, since $|z - x| \le |z - y| + |y - x| \le \hat{r} + \bar{D}$, we have $|\bar{D} - |z - x|| \le \hat{r}$. Next, obtain the same inequality with $z = x^{k+1}$ and $\hat{r} = (r^2_k - \rho^k_\phi - \rho^k_S)^{1/2}$ to get a contradiction. □

Lemma 4.2 says that for each group there is a growing ball $B(x^{k(l)+1}, \bar{D} - r_k)$ such that if x^k enters this ball then $f^k_{\rm lev} < f^*$. Hence Lemma 4.2 may be used at Step 4(iii) to detect $f^k_{\rm lev} < f^*$. Following [Dre83], one may argue that the conditions of Lemma 4.2 will be activated earlier than the usual condition $r^2_k < \rho^k_\phi + \rho^k_S$. Indeed, r_k decreases from $\bar{D}$ to zero, whereas usually $|x^{k+1} - x^{k(l)+1}| \ll \bar{D}$, e.g., if $\bar{D}$ is a generous overestimate of diam(S).

5 Simple modifications

We shall now describe some simple modifications of Algorithm 2.2.

At Step 5(iii) one may set $x^{k+1} = x^k_{\rm rec}$, i.e., each group K_l may start from the best point found so far (if $g_f(x^k_{\rm rec})$ is stored). Alternatively, as in [KAC91], x^{k+1} could be chosen arbitrarily in S, but then Step 1 would have to evaluate f and g_f at this point, leading to a

slight deterioration in efficiency estimates such as Lemma 3.5 and Corollary 3.7 (where we would have $n_f^{k_\epsilon} = k_\epsilon$).

By suppressing null steps in our notation we may express the efficiency estimates in terms of the number of f-evaluations alone (as is customary in, e.g., [NeY79]).

Theorem 5.1. *Suppose Step* 5(iii) *sets* $f_{\rm low}^k = \hat{f}_{\rm low}^k$ *and* $\rho_k = 0$ *without increasing* k. *Then the total number of* f*-evaluations always equals* k, *and the efficiency estimate* (3.1) *holds.*

Proof. For contradiction, consider the *unmodified* algoritm. At Step 0 set $n = 1$ and $\hat{z}^1 = x^1$. At Step 6 set $\hat{z}^{n+1} = x^{k+1}$ and $\Delta_z^n = \hat{\Delta}^k$, and increase n by 1. Then at Steps 2 and 6 we always have $n = k - l$ for the current values of k, l and n, and at Step 2 $n_f^k = n$. Suppose $\Delta_z^{n_\epsilon} \geq \epsilon > 0$ for some $n_\epsilon = k_\epsilon - l^{k_\epsilon}$ at Step 6. By Lemma 3.5 and (3.1b), we have $n_\epsilon \leq c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon)^2$. Hence, for any $\epsilon > 0$, (3.1a) holds with k and $\{x^j\}_{j=1}^k$ replaced by n and $\{\hat{z}^j\}_{j=1}^n$ respectively. It remains to identify $\{\hat{z}^n\}$ with the sequence $\{x^k\}$ generated when Step 5 does not increase k. □

We conclude, in particular, that by letting $\phi^k \equiv f^k$ at Step 4, we obtain the simple SPLA of (2.2) that enjoys the efficiency estimate (3.1). One must, however, be cautious in interpreting such results, because Algoritm 2.2 could loop infinitely between Steps 2 and 5.

Corollary 5.2. *Suppose* $\epsilon_{\rm opt} > 0$ *and Step* 5(iii) *sets* $f_{\rm low}^k = \hat{f}_{\rm low}^k$ *and* $\rho_k = 0$ *without increasing* k. *Then the algoritm will terminate with* $f(x_{\rm rec}^k) \leq f^* + \epsilon_{\rm opt}$ *after* $k = 1 + k_{\epsilon_{\rm opt}}$ *iterations and* f*-evaluations, where* $k_{\epsilon_{\rm opt}} \leq c_{\rm SPLA}(t_{\min}, t_{\max}, \kappa)(\bar{D}L_f/\epsilon_{\rm opt})^2$ *with* $c_{\rm SPLA}$ *given by* (3.1b). *Moreover,* (3.1) *holds for any* $\epsilon > \epsilon_{\rm opt}$.

Proof. Arguing by contradiction, use Theorem 3.8 to deduce that any loop between Steps 2 and 5 must be finite when $\epsilon_{\rm opt} > 0$, and then apply Theorem 5.1. □

As in [KAC91], let us consider setting $f_{\rm lev}^k = f_{\rm up}^k - \kappa_k \Delta^k$ at Step 2, where $\kappa_k \in [\kappa_{\min}, \kappa_{\max}]$ for some fixed $0 < \kappa_{\min} \leq \kappa_{\max} < 1$. We only require κ_k to produce $f_{\rm lev}^k \leq f_{\rm lev}^{k-1}$ if $k > k(l)+1$ (e.g., let $\kappa_k \in [\kappa_{k-1}, \kappa_{\max}]$ for such k). Then, as before, the level can increase only after the lower bound increases (i.e., (2.15) holds). Clearly, we must replace κ by $\kappa_{\max}$ in (2.16) and (2.17), and by $\kappa_{\min}$ in Lemmas 3.1 and 3.4. Similar replacements should be made in the remaining efficiency results. For instance, (3.9b) becomes

$$\hat{c}_{\rm RLA}(t_{\min}, t_{\max}, \kappa_{\min}, \kappa_{\max}) = 1/t_{\min}(2 - t_{\max})\kappa_{\min}^2(1 - \kappa_{\max}^2) - 1/2e\ln(\kappa_{\max}),$$

where again the 'best' $\kappa_{\min} = \kappa_{\max} \approx 0.677653$. We conclude that this modification cannot improve the preceding efficiency estimates. It may, however, be useful in practice to choose small κ_k at initial iterations in order to reduce the dependence on $f_{\rm low}^k$ until it is improved.

6 Using the known optimal value

Let us now consider the case when f^* is known.

Theorem 6.1. *If $f^1_{\rm low} = f^*$ then $l \equiv 0$ and the following efficiency estimate holds:*

$$k > c^*_{\rm RLA}(t_{\min}, t_{\max}, \kappa)(\mathrm{diam}(S)L_f/\epsilon)^2 \quad \Rightarrow \quad f^k_{\rm up} - f^* \le \hat{\Delta}^k < \epsilon, \tag{6.1a}$$

$$c^*_{\rm RLA}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2, \tag{6.1b}$$

$$\min c^*_{\rm RLA}(\cdot, \cdot, \kappa) = c^*_{\rm RLA}(1, 1, \kappa) = 1/\kappa^2. \tag{6.1c}$$

Moreover, one may use $\kappa = 1$ and $f^k_{\rm lev} \equiv f^*$, in which case (6.1) reduces to (1.4).

Proof. Use (2.1), (2.12) and Lemma 3.3 to deduce that Step 5 cannot be entered (i.e., $l \equiv 0$) and $f^k_{\rm lev} \ge f^k_{\rm low} = f^*$ for all k. Next, invoke (3.5) in the proof of Lemma 3.4 in order to replace $\bar{D}$ by $\mathrm{diam}(S)$ in Lemmas 3.4 and 3.5. Finally, observe that m and $(1 - \kappa^2)$ may be dropped from (3.7) to give (6.1), since $m = 0$ and $k_\epsilon \le c/\epsilon^2$ in part (ii) of the proof of Lemma 3.5, which remains valid even if $\kappa = 1$ because no summation is required. □

We conclude that if f^* is known then Step 5 and the tests of Step 4 may be omitted, so that $\bar{D}$ is not required. Moreover, setting $f^k_{\rm lev} \equiv f^*$ ($\kappa = 1$ in (6.1)) gives the 'best' efficiency estimate (1.4). In particular, (1.4) holds for for the simplest method of (1.2) (using $\phi^k \equiv f^k$ at Step 4), as well as for Polyak's accelerated method from [Pol69] (with $\phi^k \equiv \hat{f}^k$; cf. (2.5)).

Remark 6.2. Note that, by Lemma 3.3, $f^k_{\rm low} \equiv f^*$ ensures Fejér monotonicity $|x^* - x^{k+1}| \le |x^* - x^k|$ for all k and $x^* \in S^*$. Hence one easily checks that $\mathrm{diam}(S)$ and L_f in (6.1) may be replaced by $D^* = |x^* - x^1|$ and $L^*_f = \sup\{|g_f(x)| : |x^* - x| \le D^*\}$ for any $x^* \in S^*$. Thus one may get an efficiency estimate even for *unbounded* S if S^* is nonempty! Also Fejér monotonicity and Theorem 3.8 imply that $\{x^k\}$ *converges* to an optimal point (let x^* be a cluster point of $\{x^k_{\rm rec}\}$). The question whether $\{x^k\}$ converges for other level controls is left open for future research.

The same argument also shows that if we chose $f^1_{\rm low} > f^*$ then either termination would occur with $f^k_{\rm up} \le f^1_{\rm low} + \epsilon_{\rm opt}$ or (6.1) would hold with f^* replaced by $f^1_{\rm low}$ (as if f were replaced by $\max\{f, f^1_{\rm low}\}$).

7 Level control via frozen level gaps

In Algorithm 2.2 we have $f^k_{\rm up} - f^k_{\rm lev} = \kappa\Delta^k$, i.e., the desired objective reduction is a fraction of the current gap. An alternative technique consists in freezing the level gap $\Delta^k_{\rm lev} = f^k_{\rm up} - f^k_{\rm lev}$ at $\kappa\Delta^{k(l)}$ between iterations $k(l)$ and $k(l+1)$ that increase the lower bound.

Thus we modify Algorithm 2.2 as follows. Step 2 sets $f^k_{\rm lev} = f^k_{\rm up} - \Delta^k_{\rm lev}$, with $\Delta^1_{\rm lev} = \kappa\Delta^1$; Step 5 sets $\Delta^{k+1}_{\rm lev} = \kappa\hat{\Delta}^k$, whereas Step 6 sets $\Delta^{k+1}_{\rm lev} = \Delta^k_{\rm lev}$.

It is easy to check that the relations that ensure (2.14) continue to hold, whereas (2.15) follows from the fact that $f^{k+1}_{\rm up} \le f^k_{\rm up}$, while $\Delta^k_{\rm lev} = \kappa\hat{\Delta}^{k(l)}$ if $k(l) < k \le k(l+1)$ and $l \ge 0$, where $\hat{\Delta}^0 = \Delta^1$. Next, for $k = k(l+1)$ at Step 5 we have $\hat{f}^k_{\rm low} \ge f^k_{\rm lev} = f^k_{\rm up} - \Delta^k_{\rm lev}$, so

$$\hat{\Delta}^{k(l+1)} \le \Delta^k_{\rm lev} = \kappa\hat{\Delta}^{k(l)} \quad \text{if} \quad k(l) < k \le k(l+1) \text{ and } l \ge 0 \tag{7.1}$$

and (2.17) follow by induction. Notice that the algorithm may also go to Step 5 from Step 2 if $f^k_{\rm lev} \le f^k_{\rm low}$. In words: each group K_l ends by discovering that the target level is unattainable

(and possibly that the lower bound may be increased). Then the level is raised by setting the level gap to a fraction of the 'true' gap (between the bounds). The remaining level reductions within each group occur only when the objective improves, with the level gap and the lower bound staying fixed.

The efficiency analysis for the modified algorithm is similar to that for Algorithm 2.2, so we shall only indicate changes. Lemmas 3.2 and 3.3 remain valid. In Lemma 3.1 we may replace $\kappa\Delta^k$ by $\Delta^k_{\rm lev}$ (using $f(x^k) - f^k_{\rm lev} \geq f^k_{\rm up} - f^k_{\rm up} + \Delta^k_{\rm lev} = \Delta^k_{\rm lev}$), so (3.6) is replaced by

$$k - k(l) \leq (\bar{D}L_f/\Delta^k_{\rm lev})^2/t_{\min}(2 - t_{\max}) \quad \text{if} \quad k(l) < k < k(l+1) \text{ and } \Delta^k_{\rm lev} > 0. \tag{7.2}$$

In part (i) of the proof of Lemma 3.5 refer to (7.1) (instead of (2.16)) to get $\Delta^k_{\rm lev} \geq \epsilon/\kappa^{m-l-1}$ for all $k \in K_l \cap K(\epsilon)$ and $l = 0{:}m$, and use this relation and (7.2) in part (ii) to get the previous bounds. The remaining convergence results of §3 are easy to verify.

Another interesting modification is described in the following

Theorem 7.1. *If we set* $\epsilon_{\rm opt} = \Delta^1_{\rm lev} = \epsilon > 0$ *(and possibly* $f^1_{\rm low} = -\infty$*) then the modified algoritm will terminate with* $f^k_{\rm up} - f^* \leq \epsilon$ *and* $l = 0$ *at iteration* $k = 1 + k_\epsilon$*, where*

$$k_\epsilon \leq (\bar{D}L_f/\epsilon)^2/t_{\min}(2 - t_{\max}). \tag{7.3}$$

Proof. If Step 5 is not entered for $k = 1{:}k_\epsilon$ then (7.2) with $\Delta^k_{\rm lev} = \epsilon$ and $l = 0$ implies (7.3). Iteration $k = k_\epsilon + 1$ terminates at Step 2, or at Step 5 with $\hat{\Delta}^k \leq \Delta^1_{\rm lev} = \epsilon$ (cf. (7.1)). □

A result essentially equivalent to the above theorem is given in [KuF90] for the simplest case of $\phi^k \equiv f^k$ at Step 4. A comparison with all the preceding efficiency estimates (especially Corollary 5.2) suggests that, for a given accuracy $\epsilon_{\rm opt} > 0$, the strategy of Theorem 7.1 yields the best estimate. We believe, however, that in practice a 'small' $\Delta^1_{\rm lev} = \epsilon_{\rm opt}$ might result in a slow 'short-step' method, whose behavior would be close to the worst-case estimate even for 'well-behaved' objectives. On the other hand, one may set $\Delta^1_{\rm lev}$ to an estimate of $f(x^1) - f^*$ (if any), so as to exploit any extra information at initial iterations; once a 'reasonable' $f^k_{\rm low}$ is obtained then a switch to the original level strategy of Algorithm 2.2 may occur. (A similar idea is used in [LNN91].)

8 Level control via full model minimization

As in [LNN91], the *best underestimate* of f^* at iteration k is given by $\check{f}^k_{\min} = \min_S \check{f}^k$ with $\check{f}^k = \max_{j=1:k} f^j$. Let us, therefore, consider a version of Algorithm 2.2 in which Step 2 sets $f^k_{\rm low} = \check{f}^k_{\min}$, Step 4(i) chooses $\phi^k \leq \check{f}^k$, and Steps 4(iii) and 5 are deleted because $\bar{D}$ is no longer required for updating $f^k_{\rm low}$. (Note that $\mathcal{L}(\phi^k, f^k_{\rm lev}) \neq \emptyset$ at Step 4(ii) due to $f^k_{\rm lev} > \check{f}^k_{\min} \geq \inf \phi^k$ by (2.1).)

Since $\check{f}^k \leq \max\{\check{f}^k, f^{k+1}\} = \check{f}^{k+1} \leq f$, we still have $f^{k+1}_{\rm low} \geq f^k_{\rm low}$, $\Delta^{k+1} \leq \Delta^k$ and (2.14) for all k. Next,

$$\Delta^k < \kappa\Delta^j \quad \text{if} \quad \check{f}^k_{\min} > f^j_{\rm lev} \text{ and } j < k, \tag{8.1}$$

since then $\check{f}^k_{\min} > f^j_{\rm up} - \kappa\Delta^j \geq f^k_{\rm up} - \kappa\Delta^j$ by (2.1) with $\Delta^k = f^k_{\rm up} - \check{f}^k_{\min}$.

Theorem 8.1. *The following efficiency estimate holds for each $\epsilon > 0$:*

$$k > c_{\mathrm{LNN}}(t_{\min}, t_{\max}, \kappa)(\mathrm{diam}(S)L_f/\epsilon)^2 \quad \Rightarrow \quad f^k_{\mathrm{up}} - f^* \le \Delta^k < \epsilon, \tag{8.2a}$$

$$c_{\mathrm{LNN}}(t_{\min}, t_{\max}, \kappa) = 1/t_{\min}(2 - t_{\max})\kappa^2(1 - \kappa^2), \tag{8.2b}$$

$$\min c_{\mathrm{LNN}}(\cdot,\cdot,\cdot) = c_{\mathrm{LNN}}(1, 1, 1/\sqrt{2}) = 4. \tag{8.2c}$$

Proof. (i) Suppose $\Delta^{k_\epsilon} \ge \epsilon > 0$ for some k_ϵ. Let us split $K(\epsilon) = \{1\!:\!k_\epsilon\}$ into groups $\check{K}_l$, $l = 1\!:\!m$ as follows. Let $\check{k}(1) = k_\epsilon$. For $l = 1, 2, \ldots$, set $\check{K}_l = \{k \le \check{k}(l) : \Delta^k \le \Delta^{\check{k}(l)}/\kappa\}$ and $\check{k}(l+1) = \min\{k : k \in \check{K}_l\} - 1$ until $\check{k}(l+1) = 0$, and then set $m = l$. By construction, $\Delta^k \ge \epsilon/\kappa^{l-1}$ for all $k \in \check{K}_l = \{\check{k}(l+1) + 1\!:\!\check{k}(l)\}$ and $l = 1\!:\!m$.

(ii) Fix $1 \le l \le m$ and let $\check{y}^l \in \mathrm{Arg}\min_S \check{f}^{\check{k}(l)}$. By (i) and (8.1), $\check{f}^{\check{k}(l)}_{\min} \le f^k_{\mathrm{lev}}$ for all $k \in \check{K}_l$. Hence, since $\check{f}^k$ are nondecreasing and $\phi^k \le \check{f}^k$, we have $\check{y}^l \in \mathcal{L}(\phi^k, f^k_{\mathrm{lev}})$ at Step 4 for all $k \in \check{K}_l$. Therefore, $|\check{K}_l| \le (\mathrm{diam}(S)L_f/\kappa\Delta^{\check{k}(l)})^2/t_{\min}(2 - t_{\max})$ by Lemmas 3.1 and 3.2, with $k_1 = \check{k}(l+1) + 1$, $k_2 = \check{k}(l)$ and $y = \check{y}^l \in S$.

(iii) Let $c = (\mathrm{diam}(S)L_f/\kappa)^2/t_{\min}(2 - t_{\max})$. By (i) and (ii),

$$k_\epsilon = \sum_{l=1}^{m} |\check{K}_l| \le \sum_{l=1}^{m} (c/\epsilon^2)\kappa^{2(l-1)} \le c/\epsilon^2(1 - \kappa^2). \qquad \square$$

Theorem 8.1 subsumes a result in [LNN91] obtained for $\phi^k \equiv \check{f}^k + \delta_S$ and $t_k \equiv 1$. Thus it shows that the good efficiency estimate for the level method of [LNN91] comes from level control, rather than from full projection subproblems.

It is easy to verify Theorem 3.8 and Corollary 5.2 (with $\bar{D} = \mathrm{diam}(S)$) for the modified method. Moreover, we may consider setting $f^k_{\mathrm{lev}} = f^k_{\mathrm{up}} - \kappa_k\Delta^k$, where $\kappa_k \in [\kappa_{\min}, \kappa_{\max}] \subset (0, 1)$. Then (8.2) involves $\hat{c}_{\mathrm{LNN}}(t_{\min}, t_{\max}, \kappa_{\min}, \kappa_{\max}) = 1/t_{\min}(2 - t_{\max})\kappa^2_{\min}(1 - \kappa^2_{\max})$, where again the 'best' $\kappa_{\min} = \kappa_{\max} = 1/\sqrt{2}$. To check this, replace κ by $\kappa_{\max}$ in (8.1) and part (i) of the proof of Theorem 8.1, and by $\kappa_{\min}$ in part (ii).

Although it eliminates the need for $\bar{D}$, finding $\check{f}^k_{\min} = \min_S \max_{j=1:k} f^j$ may require too much storage and work per iteration when k is large. Let us, therefore, consider the following *partial model minimization strategy.* At Step 2 find $\hat{f}^k_{\min} \le \min_S \hat{f}^k$ (cf. (2.5)) and set $f^k_{\min} = \max\{\hat{f}^k_{\min}, f^{k-1}_{\min}\}$ (with $f^0_{\min} = f^1_{\mathrm{low}}$). If $f^k_{\mathrm{lev}} \le f^k_{\min}$, go to Step 5, choosing $\hat{f}^k_{\mathrm{low}} \ge f^k_{\min}$. Clearly, the efficiency results of the preceding sections remain true. Although this technique does not eliminate the dependence on $\bar{D}$ in theory, we believe that when $\hat{f}^k$ are chosen 'rich enough' (cf. §12), it will ensure better performance in practice. Also note that, to save work, $\min_S \hat{f}^k$ need not be found exactly.

9 Dual level methods

We shall now show how to use dual (ϵ-subgradient) techniques for constructing models of f that generalize those in [KuF90, LNN91]. In the simplest case such models are aggregate linearizations of f that are convex combinations of the ordinary linearizations f^j. We shall later relate them to surrogate inequalities used in relaxation methods. We start with an abstract framework that will cover several examples motivated by the the following representation

$$\partial_\epsilon \hat{f}^k(x^k) = \{ \sum_{j \in J^k} \lambda_j g_j(x^j) : \lambda_j \ge 0, j \in J^k, \sum_{j \in J^k} \lambda_j = 1, \sum_{j \in J^k} \lambda_j [\hat{f}^k(x^k) - f^j(x^k)] \le \epsilon \}. \tag{9.1}$$

Definition 9.1. *Let $\mu > 0$ be a fixed parameter of Algorithm 2.2. At Step 4 let Φ_μ^k denote the set of all closed proper convex functions $\phi^k\colon \mathbb{R}^N \to (-\infty, \infty]$ that satisfy*

$$S^* \subset \mathcal{L}(\phi^k, f_{\text{lev}}^k) \quad \text{and} \quad d_{\mathcal{L}(\phi^k, f_{\text{lev}}^k)}(x^k) \geq \mu\kappa\Delta^k/L_f \quad \text{if} \quad f_{\text{lev}}^k \geq f^*. \tag{9.2}$$

Lemma 9.2. *Let $0 < \mu_\epsilon \leq 1$ and $\mu_g > 0$ be fixed parameters of Algorithm 2.2. At Step 4 let $\Phi_{\mu_\epsilon,\mu_g}^k$ denote the set of all functions of the form $\phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k + \langle p^k, \cdot - x^k \rangle$, where $\hat{\phi}^k\colon \mathbb{R}^N \to (-\infty, \infty]$ is closed, proper and convex, $\hat{\phi}^k \in \Phi$ if $f_{\text{lev}}^k \geq f^*$, $p^k \in \partial_{\epsilon_k}\hat{\phi}^k(x^k)$ satisfies $|p^k| \leq L_f/\mu_g$, and $\epsilon_k \in [0, \epsilon_{\max}^k]$ with*

$$\epsilon_{\max}^k = \hat{\phi}^k(x^k) - f_{\text{up}}^k + (1-\mu_\epsilon)(f_{\text{up}}^k - f_{\text{lev}}^k) = \hat{\phi}^k(x^k) - f_{\text{lev}}^k - \mu_\epsilon\kappa\Delta^k. \tag{9.3}$$

Suppose Step 4(ii) uses such a ϕ^k (although it need not majorize f^k). If $f_{\text{lev}}^k \geq f^*$ then $S^* \subset \mathcal{L}(\phi^k, f_{\text{lev}}^k)$, $y^k = P_{\mathcal{L}(\phi^k, f_{\text{lev}}^k)}(x^k) = x^k - (\hat{\phi}^k(x^k) - \epsilon_k - f_{\text{lev}}^k)p^k/|p^k|^2$ and

$$\rho_\phi^k/t_{\min}(2 - t_{\max}) \geq d^2_{\mathcal{L}(\phi^k, f_{\text{lev}}^k)}(x^k) = (\hat{\phi}^k(x^k) - \epsilon_k - f_{\text{lev}}^k)^2/|p^k|^2 \geq (\mu_\epsilon\mu_g\kappa\Delta^k/L_f)^2, \tag{9.4}$$

whereas if $\mathcal{L}(\phi^k, f_{\text{lev}}^k) = \emptyset$ then $f_{\text{lev}}^k < f^*$. Moreover, $\Phi_{\mu_\epsilon,\mu_g}^k \subset \Phi_\mu^k$ if $\mu = \mu_\epsilon\mu_g$.

Proof. Suppose $f_{\text{lev}}^k \geq f^*$. Let $x \in S^*$. By construction and (2.10), $\phi^k(x) \leq \hat{\phi}^k(x) \leq f^*$ yield $S^* \subset \mathcal{L}(\phi^k, f_{\text{lev}}^k) \neq \emptyset$. By (9.3), $p^k = 0$ would imply $f^* \geq \phi^k(\cdot) = \hat{\phi}^k(x^k) - \epsilon_k \geq f_{\text{lev}}^k + \mu_\epsilon\kappa\Delta^k > f_{\text{lev}}^k$, a contradiction. Thus $y^k - x^k = -(\phi^k(x^k) - f_{\text{lev}}^k)p^k/|p^k|^2$. Since $\phi^k(x^k) - f_{\text{lev}}^k = \hat{\phi}^k(x^k) - \epsilon_k - f_{\text{lev}}^k \geq \mu_\epsilon\kappa\Delta^k$ from (9.3) and $|p^k| \leq L_f/\mu_g$ by the choice of p^k, we have (9.4), which yields (9.2) if $\mu = \mu_\epsilon\mu_g$. □

Let us consider efficiency before examples. The preceding proofs hinge on $\phi^k \in \Phi_1^k$ *only* (cf. (2.15) and the proof of Lemma 3.3), so $\phi^k \notin \Phi$ is admissible if $f_{\text{lev}}^k < f^*$ (this is used in §14). Hence Step 4 may use any $\phi^k \in \Phi_{\mu_\epsilon,\mu_g}^k \cup \Phi_\mu^k$ with $\mu = \mu_\epsilon\mu_g > 0$. Then, comparing (9.2) and (9.4) with Lemma 3.1, we see that *only the first terms* of the constants in all the preceding efficiency estimates and the right side of (7.3) need be *divided by* μ^2; of course, Δ_{lev}^k replaces $\kappa\Delta^k$ in (9.2), (9.3) and (9.4) for the frozen level gaps of §7.

As in §8, suppose $f_{\text{low}}^k \equiv \check{f}_{\min}^k$ at Step 2, Step 5 is deleted, and Step 4 chooses $\hat{\phi}^k \leq \check{f}^k$ and ϕ^k as in Lemma 9.2. Then Theorem 8.1 holds with c_{LNN} divided by $\mu^2 = \mu_\epsilon^2$, since $\mu_g = 1$ (in part (ii) of its proof use (9.4) to replace κ by $\mu\kappa$).

It should be clear that, as in §5 and §8, we may use variable $\kappa_k \in [\kappa_{\min}, \kappa_{\max}]$ and $\mu_{\epsilon,k} \in [\mu_{\min}, 1] \subset (0, 1]$ in (9.3). Then the efficiency constants are divided by $(\mu_{\min}\mu_g)^2$.

For choosing $\hat{\phi}^k$ in Lemma 9.2, note that any $\hat{\phi}^k \in \Phi$ such that $f^k \leq \hat{\phi}^k \leq \check{f}^k$ is admissible. Indeed, $f^k(x^k) = \check{f}^k(x^k) = f(x^k)$ yield $\epsilon_{\max}^k \geq 0$ in (9.3), and we may always ensure $|p^k| \leq L_f$ because $g_f(x^k) \in \partial\hat{\phi}^k(x^k)$ has $|g_f(x^k)| \leq L_f$. In view of Remark 3.9, to enlarge ρ_ϕ^k in (9.4), we may let

$$p^k = \arg\min\{\, |p|^2/2 : p \in \partial_{\epsilon_k}\hat{\phi}^k(x^k) \,\}. \tag{9.5}$$

For example, if we use $\hat{\phi}^k = \hat{f}^k$ and (9.1), then p^k may by found by QP. Since there is no need to solve (9.5) exactly, iterative QP methods (e.g., parallel relaxation-type methods) and various heuristics may by employed to save work. Note that (9.5) minimizes the denominator

in (9.4) with a fixed nominator. An alternative construction is described in the following lemma (in which one may assume $\hat{\phi}^k = \hat{f}^k$ for the first reading). It implies that we may use convex combinations of linearizations with quite arbitrary weights without destroying the preceding efficiency estimates. Possible advantages of such combinations are discussed later.

Lemma 9.3. *Let* $\hat{\phi}^k = \max_{i\in I^k} \phi_i^k$, *where* $|I^k| < \infty$, *each* ϕ_i^k *is as affine function of the form* $\phi_i^k(x) = \phi_i^k(x^k) + \langle p_{\phi_i}^k, x - x^k\rangle$ *with* $|p_{\phi_i}^k| \le L_f/\mu_g$, *and* $\phi_i^k \in \Phi$ *if* $f_{\rm lev}^k \ge f^*$. *Suppose* $\hat{\phi}^k(x^k) > f_{\rm lev}^k$, $\hat{I}^k \subset \{i \in I^k : \phi_i^k(x^k) \ge f_{\rm lev}^k\}$, $\bar{I}^k = \{i \in \hat{I}^k : \phi_i^k(x^k) = \hat{\phi}^k(x^k)\} \ne \emptyset$, $\lambda_i > 0$ *for* $i \in \hat{I}^k$, $\lambda_i = 0$ *for* $i \in I^k \setminus \hat{I}^k$, $\sum_{i\in\hat{I}^k} \lambda_i = 1$, $(p^k, \epsilon_k) = \sum_{i\in\hat{I}^k} \lambda_i (p_{\phi_i}^k, \hat{\phi}^k(x^k) - \phi_i^k(x^k))$ *and* $\phi^k = \hat{\phi}^k(x^k) - \epsilon_k + \langle p^k, \cdot - x^k\rangle$. *If* $f_{\rm lev}^k \ge f^*$ *then* $d_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k) \ge \sum_{i\in\bar{I}^k} \lambda_i(\hat{\phi}^k(x^k) - f_{\rm lev}^k)\mu_g/L_f$, *whereas if* $p^k = 0$ *then* $f_{\rm lev}^k < f^*$. *In particular,* $\phi^k \in \Phi_{\mu_\epsilon,\mu_g}^k$ *if* $\lambda_j \ge \mu_\epsilon \kappa \Delta^k/(\hat{\phi}^k(x^k) - f_{\rm lev}^k)$ *for some* $j \in \bar{I}^k$, *e.g., if* $\lambda_j \ge \mu_\epsilon$ *and* $\hat{\phi}^k(x^k) \ge f_{\rm up}^k$.

Proof. Suppose $f_{\rm lev}^k \ge f^*$, $x \in S^*$ and $j \in \bar{I}^k$. By construction, $\hat{\phi}^k(x^k) - \epsilon_k - f_{\rm lev}^k = \sum_{i\in\hat{I}^k} \lambda_i(\phi_i^k(x^k) - f_{\rm lev}^k) \ge \sum_{i\in\bar{I}^k} \lambda_i(\hat{\phi}^k(x^k) - f_{\rm lev}^k) > 0$. Hence with $\phi^k(x^k) = \hat{\phi}^k(x^k) - \epsilon_k$, $p^k = 0$ would give $f^* \ge \sum_{i\in\hat{I}^k} \lambda_i \phi_i^k(x) = \phi^k(x) = \phi^k(x^k) > f_{\rm lev}^k$, a contradiction. Thus $d_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k) = (\phi^k(x^k) - f_{\rm lev}^k)/|p^k| \ge \sum_{i\in\bar{I}^k} \lambda_i(\hat{\phi}^k(x^k) - f_{\rm lev}^k)/|p^k|$, where $|p^k| \le \sum_{i\in\bar{I}^k} \lambda_i |p_{\phi_i}^k| \le L_f/\mu_g$. Recall Lemma 9.2 and (2.1) to complete the proof. □

Supposing $f_{\rm lev}^k \ge f^*$, let us now compare the dual approach based on (9.4) (and possibly (9.5)) using $y^k = P_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k)$ with a primal one that employs $\hat{\phi}^k$ directly to find

$$\hat{y}^k = P_{\mathcal{L}(\hat{\phi}^k, f_{\rm lev}^k)}(x^k) = \arg\min\{\, |x - x^k|^2/2 : \hat{\phi}^k(x) \le f_{\rm lev}^k \,\}. \tag{9.6}$$

Lemma 9.4. *We have* $|y - \hat{y}^k|^2 \le |y - x^k|^2 - |\hat{y}^k - x^k|^2$ *and* $|y - y^k|^2 \le |y - x^k|^2 - |y^k - x^k|^2$ *for all* $y \in \mathcal{L}(\hat{\phi}^k, f_{\rm lev}^k)$, *where* $|y^k - x^k|^2 \le |\hat{y}^k - x^k|^2 - |\hat{y}^k - y^k|^2$. *Moreover,*

$$\sup\{\, (\hat{\phi}^k(x^k) - \epsilon - f_{\rm lev}^k)/|p| : \epsilon \in [0, \hat{\phi}^k(x^k) - f_{\rm lev}^k),\ p \in \partial_\epsilon \hat{\phi}^k(x^k) \,\} = |\hat{y}^k - x^k|, \tag{9.7}$$

with the supremum attained at some $\hat{\epsilon}_k$ *and* $\hat{p}^k$ *if* $\hat{\phi}^k$ *is polyhedral or* $\inf \hat{\phi}^k < f_{\rm lev}^k$.

Proof. The first assertion follows from (2.8) and $\phi^k \le \hat{\phi}^k$. Hence, recalling (9.4), $|\hat{y}^k - x^k|$ majorizes the left side of (9.7). To establish equality, suppose initially that $\hat{\phi}^k$ is polyhedral or $\inf \hat{\phi}^k < f_{\rm lev}^k$. Then, by the Kuhn-Tucker conditions for (9.6), there exist $\hat{p}^k \in \partial\hat{\phi}^k(\hat{y}^k)$ and a multiplier $\hat{\lambda} \ge 0$ such that $\hat{y}^k - x^k = -\hat{\lambda}\hat{p}^k$. Clearly, $\hat{\phi}^k(\hat{y}^k) = f_{\rm lev}^k$ and $\hat{\lambda} > 0$ because $x^k \notin \mathcal{L}(\phi^k, f_{\rm lev}^k) \supset \mathcal{L}(\hat{\phi}^k, f_{\rm lev}^k)$ $(\phi^k \le \hat{\phi}^k)$. Letting $\tilde{\phi}^k = \hat{\phi}^k(\hat{y}^k) + \langle \hat{p}^k, \cdot - \hat{y}^k\rangle$ and $\hat{\epsilon}_k = \hat{\phi}^k(x^k) - \tilde{\phi}^k(x^k) = f(x^k) - f_{\rm lev}^k + \hat{\lambda}|\hat{p}^k|^2$, we get (9.7). In the general case, replace $f_{\rm lev}^k$ in (9.6) by $t > f_{\rm lev}^k$, so that the Slater condition holds, define $\hat{y}(t)$, $\hat{p}(t)$ and $\hat{\epsilon}(t)$ as above and let $t \downarrow f_{\rm lev}^k$ with $\hat{y}(t) \to \hat{y}^k$. □

In view of Remark 3.9, the first bound of Lemma 9.4 is, in general, better than the second one. On the other hand, (9.7) says that the dual approach can, in principle, be as good as the primal one if (9.5) is used with a carefully chosen ϵ_k. Thus such bounds seem to favor the

primal approach. However, they are local and the dual one may employ inexact QP solvers, so it may be easier to implement.

Choosing $(\epsilon_k, p^k) = (\hat{\epsilon}_k, \hat{p}^k)$ to solve (9.7) gives an *'optimal' dual method* that does not need $\mu_\epsilon > 0$ in (9.3) if $\hat{\phi}^k \geq f^k$ and $\hat{\phi}^k \in \Phi$. It is, however, more difficult to implement than the equivalent primal method that may solve (9.6) via QP when $\hat{\phi}^k$ is polyhedral.

We may add that [LNN91] employs $\hat{\phi}^k = \check{f}^k$, $t_k = 1$, $f^k_{\rm lev} = \check{f}^k_{\rm min}$ and constant $\kappa, \mu_\epsilon \in (0,1)$ and $\mu_g = 1$, whereas [KuF90] proceeds as in Theorem 6.1 with $\mu_\epsilon = 1$ and $\hat{\phi}^k = f$ without specifying any models of f (but $\mu_\epsilon = 1$ may severely restrict the choice of p^k; cf. (9.3)).

10 Conditions on generalized relaxations

The following generalized version of Step 4 will allow various implementations.

Step 4' (*Generalized relaxations*).

(i) Find $z^k \in \mathbb{R}^N$ and $\rho^k_\phi \geq 0$ such that $\rho^k_\phi \geq t_{\rm min}(2 - t_{\rm max}) d^2_{\mathcal{L}(f^k, f^k_{\rm lev})}(x^k)$ and

$$|y - z^k|^2 \leq |y - x^k|^2 - \rho^k_\phi \quad \forall y \in S^* \quad \text{if} \quad f^k_{\rm lev} \geq f^*: \tag{10.1}$$

if $f^k_{\rm lev} < f^*$ then z^k and ρ^k_ϕ are arbitrary (even $\rho^k_\phi = \infty$ is admissible). If $\rho_k + \rho^k_\phi > \bar{D}^2$ or it is discovered by another test that $f^k_{\rm lev} < f^*$, go to Step 5.

(ii) Find $x^{k+1} \in S$ and $\rho^k_S \geq 0$ such that $|y - x^{k+1}|^2 \leq |y - z^k|^2 - \rho^k_S$ for all $y \in S$. If $\rho_k + \rho^k_\phi + \rho^k_S > \bar{D}^2$, go to Step 5; otherwise, go to Step 6.

Naturally, the dual methods of §9 replace f^k by ϕ^k in Step 4'(i), whereas if $f^k_{\rm lev} \equiv \check{f}^k_{\rm min}$ as in §8 then S^* in (10.1) should be replaced by $\{y \in S : \check{f}^k(y) \leq f^k_{\rm lev}\}$. It is easy to verify all the preceding efficiency results for such modifications. (Hint: let $y \in S^*$ in Lemma 3.2, with S^* replaced by $\mathcal{L}(\check{f}^k, f^k_{\rm lev})$ in §8.)

11 Using general relaxation methods

We now show how to implement Step 4' via general relaxation methods for linear inequalities; see, e.g., [AhC89, Kiw92] and the references therein.

Suppose ϕ^k is polyhedral, so that $\mathcal{L}(\phi^k, f^k_{\rm lev})$ has the form $\{x : \langle a^i, x\rangle \leq b_i, i \in I\}$. Let $C_i = \{x : \langle a^i, x\rangle \leq b_i\}$, $i \in I$. Given a starting point $\tilde{x}^1 \notin \cap_{i\in I} C_i$, many relaxation methods attempt to find a point in $\cap_i C_i$ via the iteration

$$\tilde{x}^{n+1} = \sum_{i \in I} \tilde{\lambda}^n_i \mathcal{R}_{C_i, \tilde{t}_n}(\tilde{x}^n), \quad n = 1, 2, \ldots, \tag{11.1}$$

where the weights $\tilde{\lambda}^n_i \geq 0$, $i \in I$, satisfy $\sum_i \tilde{\lambda}^n_i = 1$, and $0 < \tilde{t}_n \leq 2$. By (2.8),

$$|y - \mathcal{R}_{C_i, \tilde{t}_n}(\tilde{x}^n)|^2 \leq |y - \tilde{x}^n|^2 - \tilde{t}_n(2 - \tilde{t}_n) d^2_{C_i}(\tilde{x}^n) \quad \forall y \in C_i;$$

multiply this by $\tilde{\lambda}^n_i$, sum over i and use $\sum_i \tilde{\lambda}^n_i = 1$ and the convexity of $|\cdot|^2$ to get

$$|y - \sum_i \tilde{\lambda}^n_i \mathcal{R}_{C_i, \tilde{t}_n}(\tilde{x}^n)|^2 \leq |y - \tilde{x}^n|^2 - \tilde{t}_n(2 - \tilde{t}_n) \sum_i \tilde{\lambda}^n_i d^2_{C_i}(\tilde{x}^n) \quad \forall y \in \cap_i C_i.$$

In other words, letting $\tilde{\rho}_n = \tilde{t}_n(2-\tilde{t}_n)\sum_i \tilde{\lambda}_i^n d^2_{C_i}(\tilde{x}^n)$, we have the Fejér estimates $|y-\tilde{x}^1|^2 - |y-\tilde{x}^n|^2 \le \sum_{j=1}^{n-1}\tilde{\rho}_j \ \forall y \in \cap_i C_i$. Therefore, if we start from $\tilde{x}^1 = P_{\mathcal{L}(f^k, f^k_{\rm lev})}(x^k)$ and terminate *for any* $n \ge 1$, then $z^k = \tilde{x}^n$ and $\rho^k_\phi = d^2_{\mathcal{L}(f^k, f^k_{\rm lev})}(x^k) + \sum_{j=1}^{n-1}\tilde{\rho}_j$ will satisfy the requirements of Step 4' (in particular we may stop if $\rho_k + \rho^k_\phi > \bar{D}^2$ for such ρ^k_ϕ). Moreover, ρ^k_ϕ may be increased by using more refined Fejér estimates to replace $\tilde{\rho}_j$ with some larger quantities [Kiw92]. In fact [Kiw92] shows that other relaxation methods have much better Fejér estimates; hence they could provide more efficient implementations of Step 4'(i).

Similar ideas may be used for implementing Step 4'(ii) via finite iterative methods that do not necessarily compute x^{k+1} as the projection of z^k onto S; see [KuF90] for details.

It is worth observing that many relaxation methods are highly amenable to parallel computation; see [AhC89, Kiw92]. Since we do not require exact projections, various heuristics may limit the work spent on relaxations.

12 QP-based implementations

We shall now discuss possible implementations of our methods that employ subgradient selection and aggregation. These two techniques have proved to be highly useful in implementations of other NDO bundle methods; see, e.g., [Kiw85, Kiw89, Kiw90] for details.

First, we describe *subgradient selection*. If $\phi^k = \hat{f}^k$ and $\mathcal{L}(\phi^k, f^k_{\rm lev}) \ne \emptyset$ then

$$y^k = \arg\min\{\, |x-x^k|^2/2 : f^j(x) \le f^k_{\rm lev}, j \in J^k \,\}. \tag{12.1}$$

Denote the Lagrange multipliers of (12.1) by λ^k_j, $j \in J^k$. Let $\hat{J}^k = \{j \in J^k : \lambda^k_j > 0\}$. By the Kuhn-Tucker (K-T) conditions, if we select $J^k_s \subset J^k$ such that $\hat{J}^k \subset J^k_s$, then J^k_s may replace J^k in (12.1) without changing its solution. This suggests that only the linearizations f^j, $j \in J^k_s$, that have contributed to y^k should be retained for the next iteration. Moreover, many QP methods will automatically produce $|\hat{J}^k| \le N$. Hence we may choose $J^{k+1} = J^k_s \cup \{k+1\}$ such that $|J^{k+1}| \le N+1$. Storing the subgradients $g^j = g_f(x^j)$ for the representation $f^j = f^j(x^k) + \langle g^j, \cdot - x^k\rangle$, we do not need x^j to update $f^j(x^{k+1}) = f^j(x^k) + \langle g^j, x^{k+1}-x^k\rangle$ for $j \in J^k_s$. Thus the required storage is of order $(N+1)^2$ (plus the QP workspace).

Since subgradient selection may require excessive storage for large N, we now turn to *subgradient aggregation*, in which aggregate linearizations are produced recursively by taking convex combinations of the 'ordinary' linearizations. Suppose $\phi^k = \max\{\hat{f}^k, \psi^{k-1}\}$ for some affine $\psi^{k-1} \in \mathrm{co}\{f^j\}_{j=1}^{k-1}$ of the form $\psi^{k-1}(\cdot) = \psi^{k-1}(x^k) + \langle g^{k-1}_\psi, \cdot - x^k\rangle$ ($\psi^0 = f^1$). Let us add to (12.1) the constraint $\psi^{k-1}(x) \le f^k_{\rm lev}$ with Lagrange multiplier λ^k_ψ. Equivalently, in terms of $d^k = y^k - x^k$, $\alpha^k_j = f^k_{\rm lev} - f^j(x^k)$, $j \in J^k$, and $\alpha^k_\psi = f^k_{\rm lev} - \psi^{k-1}(x^k)$, we must find

$$d^k = \arg\min\{\, |d|^2/2 : \langle g^j, d\rangle \le \alpha^k_j, j \in J^k, \langle g^{k-1}_\psi, d\rangle \le \alpha^k_\psi \,\}. \tag{12.2}$$

Letting $\lambda^k_s = \sum_{j\in J^k}\lambda^k_j + \lambda^k_\psi$, we define 'normalized' multipliers $\hat{\lambda}^k_j = \lambda^k_j/\lambda^k_s$, $j \in J^k$, $\hat{\lambda}^k_\psi = \lambda^k_\psi/\lambda^k_s$ that form a convex combination. Then, by the K-T conditions, $d^k = -\lambda^k_s g^k_\psi$, where $g^k_\psi = \sum_{j\in J^k}\hat{\lambda}^k_j g^j + \hat{\lambda}^k_\psi g^{k-1}_\psi \in \partial\phi^k(y^k)$. (Incidentally, $\lambda^k_s > 0$ because $y^k \ne x^k$ due to $f(x^k) > f^k_{\rm lev}$.) Defining the next *aggregate linearization* $\psi^k(\cdot) = \phi^k(y^k) + \langle g^k_\psi, \cdot - y^k\rangle$, we observe that

$\psi^k = \sum_{j\in J^k} \hat\lambda_j^k f^j + \hat\lambda_\psi^k \psi^{k-1} \in \mathrm{co}\{f^j\}_{j=1}^k$ and $y^k = P_{\mathcal{L}(\psi^k, f^k_{\mathrm{lev}})}(x^k)$. In effect, ψ^k embodies all the past subgradient information that determined y^k (equivalently, this amounts to replacing the constraints of (12.2) by their convex combination with normalized multipliers). With such motivation, the next iteration may use $\phi^{k+1} = \max\{\hat f^{k+1}, \psi^k\}$ with J^{k+1} containing $k+1$ and, e.g., all but one elements of J^k to ensure bounded storage.

An alternative *selective aggregation* consists in aggregating just two linearizations. Specifically, if we pick $i, j \in J^k$ with $\lambda_i^k, \lambda_j^k > 0$, replace f^j by $(\lambda_i^k f^i + \lambda_j^k f^j)/(\lambda_i^k + \lambda_j^k)$ and drop i from J^k, then the solution of (12.1) is unchanged and the new $f^j \in \Phi$ ($f^j \le \hat f^k$). In other words, we may replace f^j by the aggregate of f^i and f^j and destroy f^i to make room for the next f^{k+1}. (Here aggregation only limits the loss of information necessary to ensure bounded storage. In other bundle methods [Kiw85], it is crucial for convergence.)

Remark 12.1. The simplest case of aggregating just two linearizations, i.e., $J^k = \{k\}$ in (12.2), may be handled analytically. Suppose g^k and g_ψ^{k-1} are independent (the other case involves projecting on one halfspace only). Then one of the following three cases may arise: $\lambda_k^k = -\alpha_k^k/|g^k|^2$ and $\lambda_\psi^k = 0$ (if $\alpha_k^k \langle g^k, g_\psi^{k-1}\rangle \le \alpha_\psi^k |g^k|^2$); $\lambda_k^k = 0$ and $\lambda_\psi^k = -\alpha_\psi^k/|g_\psi^{k-1}|^2$; or

$$\lambda_k^k = (\langle g^k, g_\psi^{k-1}\rangle \alpha_\psi^k - |g_\psi^{k-1}|^2 \alpha_k^k)/(|g^k|^2 |g_\psi^{k-1}|^2 - \langle g^k, g_\psi^{k-1}\rangle^2), \tag{12.3a}$$

$$\lambda_\psi^k = (\langle g^k, g_\psi^{k-1}\rangle \alpha_k^k - |g^k|^2 \alpha_\psi^k)/(|g^k|^2 |g_\psi^{k-1}|^2 - \langle g^k, g_\psi^{k-1}\rangle^2). \tag{12.3b}$$

In particular, if $\alpha_\psi^k = 0$ then either $d^k = -\alpha_k^k g^k/|g^k|^2$ if $\langle g^k, g_\psi^{k-1}\rangle \ge 0$, or

$$-d^k/\lambda_k^k = g^k - \langle g^k, g_\psi^{k-1}\rangle g_\psi^{k-1}/|g_\psi^{k-1}|^2 \quad \text{and} \quad \langle d^k, g_\psi^{k-1}\rangle = 0 \quad \text{if} \quad \langle g^k, g_\psi^{k-1}\rangle < 0, \tag{12.4}$$

where $g_\psi^{k-1} = -d^{k-1}/\lambda_s^{k-1}$ if $k > 1$, so that $-d^k/\lambda_k^k = g^k - \langle g^k, d^{k-1}\rangle d^{k-1}/|d^{k-1}|^2$ and $\langle d^k, d^{k-1}\rangle = 0$ if $\langle g^k, d^{k-1}\rangle > 0$. Hence subgradient aggregation is related to the conjugate subgradient techniques of [CFM75, ShU89]; see §14.

Let us now describe subgradient selection for the dual methods of §9. Let λ_j^k, $j \in J^k$, denote a solution to (9.5) using (9.1) for $\hat\phi^k = \hat f^k$. As in the primal case, if $\hat J^k \equiv \{j \in J^k : \lambda_j^k > 0\} \subset J_s^k \subset J^k$ then J_s^k may replace J^k in (9.1) without changing p^k, and we may select $J^{k+1} = J_s^k \cup \{k+1\}$. Again, many QP methods will ensure $|\hat J^k| \le N+1$, and the required storage is of order $(N+2)^2$.

Aggregation is natural in the dual methods, since they produce an aggregate linearization ϕ^k (from $\hat\phi^k$) that determines y^k. Specifically, employing $\hat\phi^k = \max\{\hat f^k, \phi^{k-1}\}$ in (9.5) to find ϕ^k, we may choose $\hat\phi^{k+1} = \max\{\hat f^{k+1}, \phi^k\}$ with J^{k+1} containing $k+1$ and all but one elements of J^k to ensure bounded storage.

The following (primal) *pairwise projections* strategy generalizes one in [KKAS87]. Having several f^j, $j \in J^k$, let $\phi^k = \max\{f^k, f^{\hat\jmath}\}$ for $\hat\jmath \in J^k$ chosen to maximize the resulting $|y^k - x^k|$ when $\{k, \hat\jmath\}$ replaces J^k in (12.1). For example, use the formula (cf. (12.3)):

$$|d^k|^2 = [(|g^k|\alpha_j^k)^2 - 2\langle g^k, g^j\rangle \alpha_j^k \alpha_k^k + (|g^j|\alpha_k^k)^2]/(|g^k|^2|g^j|^2 - \langle g^k, g^j\rangle^2) \quad \text{if} \quad \lambda_j^k, \lambda_k^k > 0.$$

Such $\hat{\jmath}$ may be included in J^{k+1}. Alternatively, if $\hat{f}^k(y^k) > f^k_{\rm lev}$, we may replace f^k by the aggregate linearization of f^k and $f^{\hat{\jmath}}$, pick $\hat{\jmath}$ such that $f^{\hat{\jmath}}(y^k) > f^k_{\rm lev}$ and recompute y^k. Of course, more than two constraints can be used at a time, and projections may continue until y^k becomes almost feasible in (12.1). Moreover, if $N \gg |J^k|$, then maintaining a matrix of inner products between g^j, $j \in J^k$, allows us to compute pairwise projections without additional expensive inner products; cf. (12.3). One may use Lemma 9.4 to show that pairwise projections are essentially equivalent to the surrogate method $S2$ of [Oko92] applied to the inequalities $f^j(x) \le f^k_{\rm lev}$, $j \in J^k$, starting from x^k. (The remaining surrogate methods of [Oko92] are obtained by using triples of inequalities and successive aggregation.)

Remark 12.2. It is worth observing that Step 4 may perform *several relaxations* using the accumulated linearizations. Specifically, at Step 4(iii), instead of going to Step 6, we may return to Step 4(i) to choose *any* $\phi^k \in \Phi$ for the next relaxation with ρ_k replaced by $\rho_k + \rho^k_\phi + \rho^k_S$ and x^k by x^{k+1} (the replacement being justified by §10); any number of such returns can occur, and all but the final one may skip the projection on S by setting $x^{k+1} = z^k$. For example, suppose J^k is so large that we do not want to solve (12.1). Then, until y^k becomes almost feasible in (12.1), each execution of Step 4 may use $\phi^k = \max\{f^{\hat{\jmath}}, \phi^{k-1}\}$, where $\hat{\jmath} \in \operatorname{Arg\,max}_{j\in J^k} f^j(x^k)$ and ϕ^{k-1} is the current aggregate linearization, i.e., it may solve (12.2) with J^k replaced by $\{\hat{\jmath}\}$. Alternatively, $\{\hat{\jmath}\}$ may be replaced by some larger set for which the solution of (12.2) is 'cheap'; cf. §13. The dual methods can be used iteratively in the same way. In other words, we may attempt to *accelerate* our algorithm by performing *extra iterations on models* of f to exploit more fully the accumulated information about f, and hence to reduce the number of f-evaluations at the cost of more work per iteration.

13 Relaxation with surrogate inequalities

This section introduces 'cheap' QP-based implementations by extending the framework of deep surrogate cuts of relaxation methods for linear inequalities [BGT81, GoT82, Tod79].

We need additional notation. For any set $\mathcal{A} \subset \mathbb{R}^N$, $\operatorname{lin}\mathcal{A}$ denotes its linear span and $\operatorname{cone}\mathcal{A} = \{a : a = \sum_{i=1}^n \lambda_i a^i, a^i \in \mathcal{A}, \lambda_i \ge 0, n < \infty\}$ denotes its convex conical hull. We let $\mathcal{A}^- = \{x : \langle x, y\rangle \le 0\ \forall y \in \mathcal{A}\}$ and $\mathcal{A}^+ = -\mathcal{A}^-$ denote its negative and positive polar cones respectively. For a matrix $A \in \mathbb{R}^{n\times n}$, a_{ij} and a^i denote its ijth element and ith column respectively. Given a set $\mathcal{I} \subset \{1{:}n\}$, $A_{\mathcal{I}}$ denotes the matrix with columns a^i, $i \in \mathcal{I}$. Matrix inequalities hold componentwisely. A is called a *Stieltjes* matrix if $a_{ij} = a_{ji} \le 0\ \forall i \ne j$, $i, j = 1{:}n$, and $A^{-1} \ge 0$.

Given $A \in \mathbb{R}^{N\times m}$ and $b \in \mathbb{R}^m$, consider the system of linear inequalities $\langle a^i, x\rangle \le b_i$, $i = 1{:}m$, having a (possibly empty) solution set $\mathcal{P} = \{x : A^T x \le b\}$. Suppose $a^i \ne 0$ for $i = 1{:}m$. Then each inequality defines a closed halfspace $H_i = \{x : \langle a^i, x\rangle \le b_i\}$, and $\mathcal{P} = \bigcap_{i=1}^m H_i$ is a convex polyhedron.

Remark 13.1. We are mainly interested in the case where $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\rm lev})$, but to compare the preceding convergence results with those for relaxation methods [Gof81, Tel82] one may observe the following. If $\mathcal{P} \ne \emptyset$ then $\mathcal{P} = \operatorname{Arg\,min} f$, where $f = \max_{i=1:m}(\langle a^i, \cdot\rangle - b_i)_+$ has a subgradient $g_f(x) = a^i$ if $\langle a^i, x\rangle - b_i = f(x) > 0$, $g_f(x) = 0$ if $f(x) = 0$, satisfying

$|g_f(x)| \le L_f := \max_{i=1:m} |a^i|$. In this case we may let $f^k_{\rm lev} \equiv f^* = 0$ in Algorithm 2.2 and proceed as if S were $\mathbb{R}^N$, replacing $\mathrm{diam}(S)$ in (1.4) by $|x^* - x^1|$ for any $x^* \in \mathcal{P}$; cf. §6. Then the SPA of (1.2) describes the *maximal residual* version of the relaxation method; the *maximal distance* version corresponds to dividing each a^i and b_i by $|a^i|$ initially.

In classical versions of relaxation and ellipsoid methods for finding a point in $\mathcal{P}$, given a current point $\tilde{x} \notin \mathcal{P}$, one finds the next point $\bar{x}$ by projecting $\tilde{x}$ on the halfspace H_i that is furthest from $\tilde{x}$, since for faster convergence one wishes to maximize $|\bar{x} - \tilde{x}|$. By combining inequalities one can sometimes obtain halfspaces that are further from $\tilde{x}$.

If $\lambda \in \mathbb{R}^m_+$, $a^\lambda = A\lambda$ and $b_\lambda = b^T\lambda$ then the *surrogate inequality* $\langle a^\lambda, x\rangle \le b_\lambda$ is valid ($A^T x \le b \Rightarrow \lambda^T A^T x \le \lambda^T b$). The *deepest* surrogate inequality that maximizes the distance $(\langle a^\lambda, \tilde{x}\rangle - b_\lambda)_+/|a^\lambda|$ from $\tilde{x}$ to $H_\lambda = \{x : \langle a^\lambda, x\rangle \le b_\lambda\}$ corresponds to

$$\tilde{\lambda} \in \mathrm{Arg\,max}\{\, \tilde{s}^T\lambda/|A\lambda| : \lambda \ge 0\,\}, \tag{13.1}$$

where $\tilde{s} := A^T\tilde{x} - b \not\le 0$ ($\tilde{x} \notin \mathcal{P}$). Clearly, if $\mathcal{P} \ne \emptyset$ then $H_{\tilde{\lambda}}$ is the unique halfspace containing $\mathcal{P}$ that is furthest from $\tilde{x}$, and $H_{\tilde{\lambda}} = \{x : \langle \tilde{d}, x - \tilde{x}\rangle \ge |\tilde{d}|^2\}$. where $\tilde{d} = P_{\mathcal{P}}(\tilde{x}) - \tilde{x}$ (since for any halfspace $H \ni P_{\mathcal{P}}(\tilde{x})$, $d_H(\tilde{x}) < d_{\mathcal{P}}(\tilde{x})$ unless $P_H(\tilde{x}) = P_{\mathcal{P}}(\tilde{x})$). Of course, $\tilde{d}$ solves the QP problem

$$\tilde{d} = \arg\min\{\, |d|^2/2 : A^T d \le -\tilde{s}\,\}. \tag{13.2}$$

By duality, we may equivalently find its (possibly nonunique) Lagrange multiplier vector

$$\tilde{\lambda} \in \mathrm{Arg\,min}\{\, |A\lambda|^2/2 - \tilde{s}^T\lambda : \lambda \ge 0\,\}, \tag{13.3}$$

Indeed, by the K-T conditions, $\tilde{d}$ and $\tilde{\lambda}$ satisfy (13.2)-(13.3) iff $\tilde{d} = -A\tilde{\lambda}$, $A^T\tilde{d} \le -\tilde{s}$, $\tilde{\lambda} \ge 0$ and $\tilde{\lambda}^T(A^T\tilde{d} + \tilde{s}) = 0$. Hence $\tilde{s}^T\tilde{\lambda} = |\tilde{d}|^2$ and $\langle a^{\tilde{\lambda}}, \tilde{x}\rangle - |\tilde{d}|^2 = \tilde{\lambda}^T A^T\tilde{x} - \tilde{x}^T A\tilde{\lambda} + b^T\tilde{\lambda} = b_{\tilde{\lambda}}$, so $\langle a^{\tilde{\lambda}}, x\rangle \le b_{\tilde{\lambda}}$ iff $\langle \tilde{d}, x - \tilde{x}\rangle \ge |\tilde{d}|^2$, and $P_{H_{\tilde{\lambda}}}(\tilde{x}) = P_{\mathcal{P}}(\tilde{x})$. (We may add that the optimal values in (13.1) and (13.3) are infinite iff $\mathcal{P} = \emptyset$. Since the objective of (13.1) is positively homogeneous, the deepest cut can also be found by solving $\min\{|A\lambda| : \tilde{s}^T\lambda = 1, \lambda \ge 0\}$, $\max\{\tilde{s}^T\lambda : |A\lambda| = 1, \lambda \ge 0\}$ or $\min\{|A\lambda|/\tilde{s}^T\lambda : \sum_i \lambda_i = 1, \lambda \ge 0\}$. We note that (13.1) is a special case of (9.7), whereas the restricted variant (9.5) does not seem to have been considered explicitly in the context of linear inequalities.)

Of course, finding the deepest surrogate inequality via (13.1)-(13.3) may be too expensive, *except when A is orthonormal*, in which case $\tilde{d} = -A\tilde{\lambda}$ and $\tilde{\lambda} = \tilde{s}$ by (13.3). In the general case, we may project on a surrogate $\tilde{\mathcal{P}} = \{x : Q^T x \le c\}$ of $\mathcal{P}$, where $Q^T x \le c$ is a surrogate of $A^T x \le b$ (so that $\mathcal{P} \subset \tilde{\mathcal{P}}$) and Q is orthonormal. As in [BGT81, GoT82, Tod79], it is convenient to work with a subset of inequalities, indexed by $\mathcal{I} \subset \{1{:}m\}$ say, that satisfy the *obtuse angles condition* $\langle a^i, a^j\rangle \le 0$ $\forall i \ne j$, $i, j \in \mathcal{I}$. Taking $\mathcal{I} = \{1{:}m\}$ first for simplicity, we now show how to construct suitable surrogates via orthogonalization (see Figure 13.1).

Lemma 13.2. *Let $\mathcal{A} = \{a^i\}_{i=1}^m$, $C = \mathrm{cone}\,\mathcal{A}$, $\hat{m} = \mathrm{rank}\,A$ and $\mathcal{G} = A^TA$. If $\langle a^i, a^j\rangle \le 0$ $\forall i \ne j$, $i, j = 1{:}m$, then:*

(i) *C contains an orthonormal system $\mathcal{Q} = \{q^i\}_{i=1}^{\hat{m}}$ such that $\mathrm{lin}\,C = \mathrm{lin}\,\mathcal{Q}$ and $A = QR$, where $Q \in \mathbb{R}^{N\times\hat{m}}$ is orthonormal and $R \in \mathbb{R}^{\hat{m}\times m}$ is upper triangular, with $r_{ii} \ge 0$*

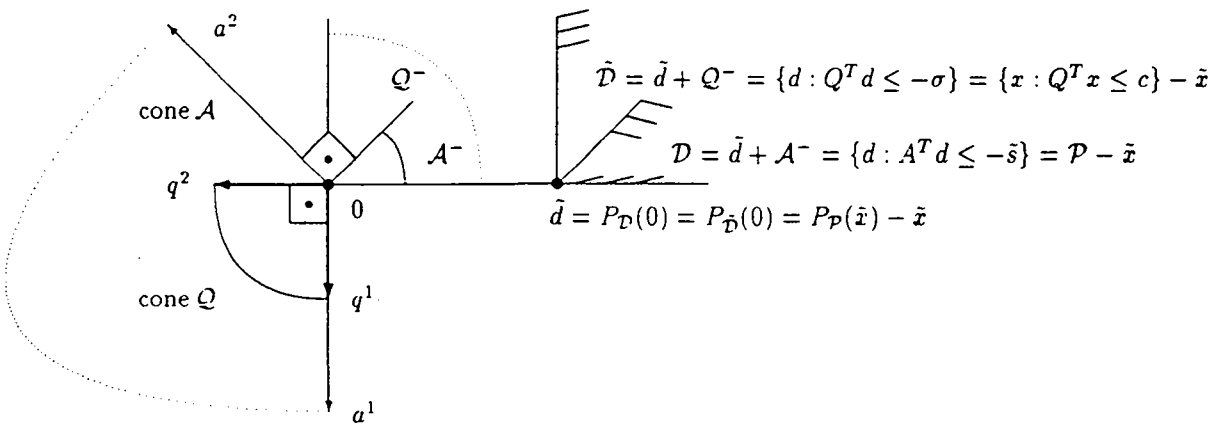

Figure 13.1: Illustration of orthogonal surrogates.

and $r_{ij} \le 0$ *for* $i = 1{:}\hat{m}$, $j = i+1{:}m$. *Q and R can be found via the Gram-Schmidt orthogonalization: set* $m_1 = 0$, *and for* $j = 1{:}m$ *set* $\tilde{q}^j = a^j - \sum_{i=1}^{m_j} \langle a^j, q^i \rangle q^i$, $r_{jj} = |\tilde{q}^j|$, $r_{ij} = \langle a^j, q^i \rangle$ *for* $i = 1{:}m_j$, $r_{ij} = 0$ *for* $i > m_j$, $i \ne j$, $q^j = \tilde{q}^j/|\tilde{q}^j|$ *and* $m_{j+1} = m_j + 1$ *if* $\tilde{q}^j \ne 0$, $m_{j+1} = m_j$ *otherwise. (The final* $m_{m+1} = \hat{m}$.)

(ii) $C^+ \cap \operatorname{lin} C \subset \mathcal{Q}^+ \cap \operatorname{lin} \mathcal{Q} = \mathcal{Q}^+ \cap \operatorname{cone} \mathcal{Q} = \operatorname{cone} \mathcal{Q} \subset C$.

(iii) *If* $\operatorname{rank} A = m$ *then* $R^{-1} \ge 0$ *and* $\mathcal{G}^{-1} \ge 0$, *i.e., $\mathcal{G}$ is a Stjeltjes matrix.*

(iv) *If* $\operatorname{rank} A = m$ *then R is the unique Cholesky factor of $\mathcal{G}$ having a positive diagonal,* $Q = AR^{-1}$ *and* $\mathcal{P} \subset \{x : Q^T x \le c\}$, *where* $c = R^{-T} b$. *In particular, each inequality* $\langle q^j, x \rangle \le c_j$ *is a surrogate of the system* $\langle q^i, x \rangle \le c_i$, $i = 1{:}j-1$, $\langle a^j, x \rangle \le b_j$, *and hence a surrogate of the (possibly stronger) system* $\langle a^i, x \rangle \le b_i$, $i = 1{:}j$.

Proof. (i) If $\mathcal{Q}_j = \{q^i\}_{i=1}^{m_j} \subset C_j = \operatorname{cone}\{a^i\}_{i=1}^{j-1}$ then $a^j \in C_j^-$ yields $\langle a^j, q^i \rangle \le 0$ for $i = 1{:}m_j$, so $\tilde{q}^j \in \operatorname{cone}(\{a^j\} \cup \mathcal{Q}_j) \subset C_{j+1}$ and $\mathcal{Q}_{j+1} \subset C_{j+1}$. The rest follows by induction.

(ii) $a \in \mathcal{Q}^+ \cap \operatorname{lin} \mathcal{Q} \iff a = \sum_{i=1}^{\hat{m}} \langle a, q^i \rangle q^i$ with $\langle a, q^i \rangle \ge 0$ for $i = 1{:}\hat{m} \iff a \in \operatorname{cone} \mathcal{Q}$, since $a = \sum_{i=1}^{\hat{m}} \lambda_i q^i$ iff $\lambda_i = \langle a, q^i \rangle$ for $i = 1{:}\hat{m}$. Clearly, $\mathcal{Q} \subset \operatorname{cone} \mathcal{Q} \subset C$ and $C^+ \subset \mathcal{Q}^+$.

(iii) If $\hat{m} = m$ then $r_{ii} > 0$ for $i = 1{:}m$, so R and $\mathcal{G} = R^T R$ are nonsingular. Since $r_{ij} \le 0$ for $i < j$, if $Ru = v \ge 0$ then $u \ge 0$ from $u_i = (v_i - \sum_{j=i+1}^{m} r_{ij} u_j)/r_{ii}$, $i = m, \ldots, 1$. Hence $R^{-1} \ge 0$ and $\mathcal{G}^{-1} = R^{-1} R^{-T} \ge 0$.

(iv) The uniqueness of R is well-known. Use $(Q^T, c) = R^{-T}(A^T, b)$ and $(q^j, c_j) = [(a^j, b_j) - \sum_{i=1}^{j-1} r_{ij}(q^i, c_i)]/r_{jj}$, $j = 1, \ldots, m$, with $R^{-T} \ge 0$ and $r_{ij} \le 0$ by (i) and (iii), to get the desired conclusion, noting that $A^T x \le b \Rightarrow R^{-T} A^T x = Q^T x \le R^{-T} b = c$. □

The next result helps in selecting subsets of inequalities for which the projections are 'easy'. For any $\mathcal{I} \subset \{1{:}m\}$, let $\mathcal{A}_\mathcal{I} = \{a^i\}_{i \in \mathcal{I}}$, $\mathcal{P}_\mathcal{I} = \{x : A_\mathcal{I}^T x \le b_\mathcal{I}\}$, $\mathcal{D}_\mathcal{I} = \{d : A_\mathcal{I}^T d \le -\tilde{s}_\mathcal{I}\}$ and $\mathcal{G}_{\mathcal{I}\mathcal{I}} = A_\mathcal{I}^T A_\mathcal{I}$. If $\mathcal{P}_\mathcal{I} \ne \emptyset$, let $\tilde{d}_{(\mathcal{I})} = P_{\mathcal{D}_\mathcal{I}}(0)$, so that $\tilde{d}_{(\mathcal{I})} = P_{\mathcal{P}_\mathcal{I}}(\tilde{x}) - \tilde{x}$ from $\tilde{s}_\mathcal{I} = A_\mathcal{I}^T \tilde{x} - b_\mathcal{I}$.

Lemma 13.3. *Suppose* $\mathcal{I} \subset \{1{:}m\}$, $\langle a^i, a^j \rangle \le 0$ $\forall i \ne j$, $i, j \in \mathcal{I}$, *and* $\tilde{s}_\mathcal{I} \ge 0$. *Then:*

(i) *If* $\operatorname{rank} A_\mathcal{I} = |\mathcal{I}|$ *then* $\tilde{d}_{(\mathcal{I})} = -A_\mathcal{I} \tilde{\lambda}_\mathcal{I}$, *where* $\tilde{\lambda}_\mathcal{I} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \tilde{s}_\mathcal{I} \ge 0$, *and*

$$\tilde{d}_{(\mathcal{I})} = \arg\min\{\, |d|^2/2 : A_\mathcal{I}^T d = -\tilde{s}_\mathcal{I} \,\}. \tag{13.4}$$

Moreover, for $\hat{\mathcal{I}} = \{i : \tilde{\lambda}_i > 0\}$ *and each* $j \in \mathcal{I}$, $\tilde{\lambda}_j > 0$ *if* $\tilde{s}_j > 0$, *if* $\tilde{s}_j = \tilde{\lambda}_j = 0$ *then* $a^j \perp \mathcal{A}_{\hat{\mathcal{I}}}$, *and* $\tilde{\lambda}_j = 0$ *if* $\tilde{s}_j = 0$ *and* $a^j \perp \mathcal{A}_{\mathcal{I}\setminus\{j\}}$.

(ii) *If* $\operatorname{rank} A_{\mathcal{I}} = |\mathcal{I}|$, $j \in \{1{:}m\} \setminus \mathcal{I}$, $A_{\mathcal{I}}^T a^j \leq 0$, $\mathcal{J} = \mathcal{I} \cup \{j\}$ *and either* $\tilde{s}_j > 0$, *or* $\tilde{s}_j = 0$ *and* $\tilde{s}_{\mathcal{I}} > 0$, *then* $\operatorname{rank} A_{\mathcal{J}} = |\mathcal{J}| \iff \mathcal{D}_{\mathcal{J}} \neq \emptyset \iff \mathcal{P}_{\mathcal{J}} \neq \emptyset$.

(iii) $\operatorname{rank} A_{\mathcal{I}} = |\mathcal{I}| \iff A_{\mathcal{I}}^T \bar{d} < 0$ *for some* $\bar{d}$.

(iv) *If* $\tilde{s}_{\mathcal{I}} > 0$ *and* $\mathcal{P}_{\mathcal{I}} \neq \emptyset$ *then* $\operatorname{rank} A_{\mathcal{I}} = |\mathcal{I}|$ *and* $\tilde{\lambda}_{\mathcal{I}} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \tilde{s}_{\mathcal{I}} > 0$.

(v) *If* $\mathcal{G}_{\mathcal{I}\mathcal{I}} = R^T R$ *is nonsingular, where* $R \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ *is upper triangular,* $Q = A_{\mathcal{I}} R^{-1}$ *and* $\sigma = R^{-T} \tilde{s}_{\mathcal{I}}$, *then* $\tilde{d}_{(\mathcal{I})} = -Q\sigma$ *and* $\tilde{\lambda}_{\mathcal{I}} = R^{-1}\sigma$ *in* (i). *Moreover, if* $r_{ii} > 0$ *for* $i = 1{:}|\mathcal{I}|$ *and* $\tilde{\mathcal{D}} = \{d : Q^T d \leq -\sigma\}$ *then* $\tilde{\mathcal{D}} \supset \mathcal{D}_{\mathcal{I}}$, $\tilde{d}_{(\mathcal{I})} = P_{\tilde{\mathcal{D}}}(0)$ *and* $Q^T \tilde{d}_{(\mathcal{I})} = -\sigma$.

Proof. (i) Replace $\{1{:}m\}$ by $\mathcal{I}$ in Lemma 13.2 to get $\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \geq 0$. Hence $\tilde{\lambda}_{\mathcal{I}} = \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \tilde{s}_{\mathcal{I}} \geq 0$ and letting $\bar{d} = -A_{\mathcal{I}} \tilde{\lambda}_{\mathcal{I}}$ we have $A_{\mathcal{I}}^T \bar{d} = -\tilde{s}_{\mathcal{I}}$, so (13.4) holds with $\bar{d} = \tilde{d}_{(\mathcal{I})}$ by the K-T conditions. Since $\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \geq 0$ is positive definite, it has a positive diagonal, so $\tilde{\lambda}_j > 0$ if $\tilde{s}_j > 0$. Next, suppose $\tilde{s}_j = 0$ and let $\mathcal{J} = \mathcal{I} \setminus \{j\}$. If $\tilde{\lambda}_j = 0$ then $0 = -\tilde{s}_j = \left\langle a^j, \bar{d} \right\rangle = -\sum_{i\in\hat{\mathcal{I}}} \tilde{\lambda}_i \langle a^j, a^i \rangle$ with $\tilde{\lambda}_i > 0$ and $\langle a^j, a^i \rangle \leq 0$ imply $a^j \perp \mathcal{A}_{\hat{\mathcal{I}}}$. Conversly, if $a^j \perp \mathcal{A}_{\mathcal{J}}$ then after symmetric permutations we have $\mathcal{G}_{\mathcal{I}\mathcal{I}} = \begin{bmatrix} \mathcal{G}_{\mathcal{J}\mathcal{J}} & 0 \\ 0 & |a^j|^2 \end{bmatrix}$ and $\mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} = \begin{bmatrix} \mathcal{G}_{\mathcal{J}\mathcal{J}}^{-1} & 0 \\ 0 & |a^j|^{-2} \end{bmatrix}$, so $\tilde{\lambda}_j = \tilde{s}_j / |a^j|^2 = 0$.

(ii) If $\operatorname{rank} A_{\mathcal{J}} = |\mathcal{J}|$ then $-A_{\mathcal{J}} \mathcal{G}_{\mathcal{J}\mathcal{J}}^{-1} \tilde{s}_{\mathcal{J}} \in \mathcal{D}_{\mathcal{J}}$. If $\operatorname{rank} A_{\mathcal{J}} = |\mathcal{I}|$ then $a^j \in \operatorname{lin} \mathcal{A}_{\mathcal{I}}$. Since $a^j \in (\operatorname{cone} \mathcal{A}_{\mathcal{I}})^-$, Lemma 13.2(ii) applied to $\mathcal{A}_{\mathcal{I}}$ yields $-a^j \in \operatorname{cone} \mathcal{A}_{\mathcal{I}}$, i.e., there exists $\lambda_{\mathcal{I}} \geq 0$ such that $-a^j = A_{\mathcal{I}} \lambda_{\mathcal{I}}$. Thus if $A_{\mathcal{I}}^T d \leq -\tilde{s}_{\mathcal{I}}$ for some d then $\langle a^j, d \rangle = -\lambda_{\mathcal{I}}^T A_{\mathcal{I}}^T d \geq \lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} \geq 0$ from $\tilde{s} \geq 0$, so $D_{\mathcal{J}} = \emptyset$ because $\lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} \geq 0 > -\tilde{s}_j$ if $\tilde{s}_j > 0$, whereas $\lambda_{\mathcal{I}}^T \tilde{s}_{\mathcal{I}} > 0 = -\tilde{s}_j$ if $\tilde{s}_j = 0$ and $\tilde{s}_{\mathcal{I}} > 0$, with $\lambda_{\mathcal{I}} \neq 0$ due to $a^j \neq 0$.

(iii) First choose $\tilde{s}_{\mathcal{I}} > 0$ and $\bar{d} = -A_{\mathcal{I}} \mathcal{G}_{\mathcal{I}\mathcal{I}}^{-1} \tilde{s}_{\mathcal{I}}$, then $\tilde{s}_{\mathcal{I}} = -A_{\mathcal{I}}^T \bar{d}$, use (ii) and induction.

(iv) Combine (i) and (iii).

(v) Clearly, $\tilde{d}_{(\mathcal{I})} = -A_{\mathcal{I}} \tilde{\lambda}_{\mathcal{I}} = -QR(R^T R)^{-1} \tilde{s}_{\mathcal{I}} = -Q\sigma$. Suppose $r_{ii} > 0$ for $i = 1{:}|\mathcal{I}|$. Apply Lemma 13.2(iii,iv) to $A_{\mathcal{I}}$ and $c = R^{-T} b_{\mathcal{I}}$ to get $\mathcal{D}_{\mathcal{I}} \subset \tilde{\mathcal{D}}$ from $\sigma = R^{-T}(A_{\mathcal{I}}^T \tilde{x} - b_{\mathcal{I}}) = Q^T \tilde{x} - c$, with $\sigma = R^{-T} \tilde{s}_{\mathcal{I}} \geq 0$ since $R^{-1} \geq 0$ and $\tilde{s}_{\mathcal{I}} \geq 0$. Hence, replacing $(A_{\mathcal{I}}, \tilde{s}_{\mathcal{I}})$ by (Q, σ) in (i), we have $P_{\tilde{\mathcal{D}}}(0) = -Q(Q^T Q)^{-1}\sigma = -Q\sigma = \tilde{d}_{(\mathcal{I})}$ and $Q^T \tilde{d}_{(\mathcal{I})} = -\sigma$. □

Lemmas 13.2–13.3 extend some results of [Tod79] in a way that is useful for algorithmic developments. For example, consider the following extension of the simultaneous projections method of [Tod79] for solving a possibly inconsistent system $A^T x \leq b$.

Procedure 13.4 (for finding a point in $\mathcal{P} = \{x : A^T x \leq b\}$).
Step 0 (*Initialization*). Select $\tilde{x}^1 \in \mathbb{R}^N$, a feasibility tolerance $\epsilon_{\text{tol}} \geq 0$ and $\tilde{D} < \infty$ such that $d_{\mathcal{P}}(\tilde{x}^1) \leq \tilde{D}$ if $\mathcal{P} \neq \emptyset$. Choose $I^0 \subset \{1{:}m\}$ such that $\operatorname{rank} A_{I^0} = |I^0|$ and $\langle a^i, a^j \rangle \leq 0$ $\forall i \neq j$, $i, j \in I^0$, e.g., $I^0 = \emptyset$. Set $\tilde{\rho}_1 = 0$ and $n = 1$.

Step 1 (*Constraint evaluation*). Calculate $s^n = A^T \tilde{x}^n - b$ and i_n such that $s_{i_n}^n = \max_i s_i^n$.

Step 2 (*Stopping criterion*). If $s_{i_n}^n \leq \epsilon_{\text{tol}}$, terminate.

Step 3 (*Selection*). Set $\tilde{I}^{n-1} = \{i \in I^{n-1} : \langle a^{i_n}, a^i \rangle \leq 0, s_i^n \geq 0\}$ and $I^n = \tilde{I}^{n-1} \cup \{i_n\}$. If desired, repeat the following for some $i \in \{1{:}m\} \setminus I^n$: if $A_{I^n}^T a^i \leq 0$ and either $s_i^n > 0$, or $s_i^n = 0$ and $A_{I^n}^T a^i \neq 0$ and either $s_{I^n}^n > 0$ or $\operatorname{rank} A_{I^n \cup \{i\}} = |I^n| + 1$, then augment I^n with i.

Step 4 (*Relaxation*). Print "$\mathcal{P} = \emptyset$" and terminate if $\operatorname{rank} A_{I^n} < |I^n|$. Otherwise set $\tilde{y}^n = P_{\mathcal{P}_{I^n}}(\tilde{x}^n) = \tilde{x}^n - A_{I^n}\lambda^n_{I^n}$ with $\lambda^n_{I^n} = \mathcal{G}^{-1}_{I^nI^n} s^n_{I^n}$. Choose a stepsize $\tilde{t}_n \in T$ and set $\tilde{x}^{n+1} = \tilde{x}^n + \tilde{t}_n(\tilde{y}^n - \tilde{x}^n)$ and $\tilde{\rho}_{n+1} = \tilde{\rho}_n + \tilde{t}_n(2 - \tilde{t}_n)|\tilde{y}^n - \tilde{x}^n|^2$.

Step 5 (*Infeasibility detection*). If $\tilde{\rho}_{n+1} > \tilde{D}^2$ or $(\tilde{D} - |\tilde{x}^{n+1} - \tilde{x}^1|)^2 > \tilde{D}^2 - \tilde{\rho}_{n+1}$, print "$\mathcal{P} = \emptyset$" and terminate.

Step 6. Increase n by 1 and go to Step 1.

If $\mathcal{P} \neq \emptyset$, we may identify Procedure 13.4 with a version of Algorithm 2.2 that minimizes $f = \max_{i=1:m}(\langle a^i, \cdot\rangle - b_i)_+$ using $f^k_{\text{lev}} \equiv f^* = 0$; cf. §6 and Remark 13.1. In particular $\mathcal{P} \subset \mathcal{P}_{I^n} \subset \{x : \langle a^{i_n}, x\rangle \leq b_{i_n}\}$ corresponds to $\phi^k \in \Phi^k_1$, and Step 4 may be validated by applying Lemma 13.3 inductively at Step 3 to get $\operatorname{rank} A_{I^n} = |I^n|$ if $\mathcal{P} \neq \emptyset$. Hence if $\mathcal{P} \neq \emptyset$ then Procedure 13.4 shares the convergence properties of Algorithm 2.2 from §6, as well as those of classical relaxation methods [Agm54, Gof81, MoS54, Tel82, Tod79], such as linear rate of convergence and possible finite termination. The infeasibility test of Step 5 is justified similarly as for Algorithm 2.2; cf. §4. Note that Step 3 may include in I^n several i with $s^n_i = 0$, e.g., $i \in I^{n-1}$ if $\tilde{t}_{n-1} = 1$ and $s^n_{I^{n-1}} = 0$ from $\tilde{x}^n = \tilde{y}^{n-1}$. It is natural to choose I^n as large as possible, although one need not insist on maximality. Of course, in practice detecting $\operatorname{rank} A_{I^n} < |I^n|$ will require tolerances tuned to the factorization of A_{I^n}.

Remark 13.5. By using the Gram matrix $\mathcal{G} = A^TA$ one may avoid expensive scalar products in updating s^n without forming $\tilde{x}^n$; cf. [Tod79]. Specifically, let $s^1 = A^T\tilde{x}^1 - b$, $\nu^1 = 0 \in \mathbb{R}^m$, $s^{n+1}_{I^n} = (1 - \tilde{t}_n)s^n_{I^n}$, $s^{n+1}_{I^n_c} = s^n_{I^n_c} - \tilde{t}_n\mathcal{G}_{I^n_cI^n}\lambda^n_{I^n}$ and $\nu^{n+1} = \nu^n + \tilde{t}_n\lambda^n$ with $\lambda^n_{I^n_c} = 0$ and $I^n_c = \{1{:}m\} \setminus I^n$ for all n, so that $\tilde{x}^n = \hat{x}^1 - A\nu^n$ and $\nu^n = \sum_{j=1}^n \tilde{t}_j\lambda^j$ for all n.

Remark 13.6. If we compute the Cholesky factorization $A^T_{I^n}A_{I^n} = R^TR$ then λ^n can be found by solving the two systems $R^T\sigma = s^n_{I^n}$ and $R\lambda^n_{I^n} = \sigma$; cf. Lemma 13.3(v). To save work, R and σ may be updated when I^n changes. However, as with normal equations for least-squares problems, one may need to employ iterative refinement to improve accuracy in the presence of rounding errors. Alternatively, one may use any stable method for computing the 'skinny' QR-factorization $A_{I^n} = QR$, where Q is orthonormal, so that $A^T_{I^n}A_{I^n} = R^TR$. The classical Gram-Schmidt process may fail due to rounding errors, but reorthogonalization can ensure higher accuracy. Moreover, by Lemma 13.3, $\tilde{d}^n = \tilde{y}^n - \tilde{x}^n$ satisfies

$$\tilde{d}^n = \arg\min\{\, |d|^2/2 : A^T_{I^n}d = -s^n_{I^n}\,\}, \tag{13.5}$$

and this equality QP problem can be solved via many well-known methods. All these aspects are treated in depth in, e.g., [Bjö90, Fle87, GMW91, GVL89].

Remark 13.7. Suppose Step 3 of Procedure 13.4 chooses $I^n = \tilde{I}^{n-1} \cup \{i_n\}$ with $s^n_{\tilde{I}^{n-1}} = 0$ (recall that $s^n_{I^{n-1}} = 0$ if $\tilde{t}_{n-1} = 1$ and $\tilde{x}^n = \tilde{y}^{n-1}$). Let $\hat{m} = |I^n|$ and let $e^{\hat{m}}$ denote column $\hat{m}$ of the $\hat{m} \times \hat{m}$ identity matrix, so that $s^n_{I^n} = s^n_{i_n}e^{\hat{m}}$. Then, using $R^T\sigma = s^n_{I^n}$ and $R\lambda^n_{I^n} = \sigma$ as in Remark 13.6, we have $\sigma = s^n_{i_n}e^{\hat{m}}/r_{\hat{m}\hat{m}}$ and only the system $R\lambda^n_{I^n} = s^n_{i_n}e^{\hat{m}}/r_{\hat{m}\hat{m}}$ must be solved. This system may be used even if $s^n_{I^{n-1}} \neq 0$. Specifically, decreasing $s^n_{I^n}$ to $\tilde{s}^n_{I^n}$ with $\tilde{s}^n_{I^{n-1}} = 0$ and $\tilde{s}^n_{i_n} = s^n_{i_n}$ corresponds to setting $\tilde{y}^n = P_{\mathcal{P}^n}(\tilde{x}^n)$, where $\mathcal{P}^n = \{x : A^T_{I^n}x \leq b_{I^n} + s^n_{I^n} - \tilde{s}^n_{I^n}\}$ satisfies $\mathcal{P}_{I^n} \subset \mathcal{P}^n \subset \{x : \langle a^{i_n}, x\rangle \leq b_{i_n}\}$, so the efficiency results remain true. This

simplification is used in [Ceg92] when $\hat{t}^{n-1} < 1$. It may, however, result in slower convergence, since $\mathcal{P}^n$ can be much bigger than $\mathcal{P}_{I^n}$. (The method of [Ceg92] scales the constraints of (13.5) before computing R, but $\tilde{d}^n$ is not affected.)

Remark 13.8. Using $\tilde{d}^n = -A_{I^n}\lambda^n_{I^n}$ and $R\lambda^n_{I^n} = s^n_{i_n}\epsilon^{\hat{m}}/r_{\hat{m}\hat{m}}$ as above, for the QR-factorization $A_{I^n} = QR$ we have $\tilde{d}^n = -s^n_{i_n}q^{\hat{m}}/r_{\hat{m}\hat{m}}$, where we may take $r_{\hat{m}\hat{m}} = |q|$ and $q^{\hat{m}} = q/|q|$ for $q = a^{i_n} - \sum_{i=1}^{\hat{m}-1}\langle a^{i_n}, q^i\rangle q^i$ as in Lemma 13.2. Thus only Q could be updated by computing some elements of $R = Q^T A_{I^n}$. However, using Q instead of R would require more storage and work if $n \leq N$, and could be less accurate without reorthogonalization.

We may add that the idea of using the obtuse angle property to identify cheap projections has wider implications. (A *cheap* projection requires only the solution of one or two linear systems in contrast to combinatorial QP.) For instance, it may be employed to accelerate general projection methods for convex feasibility problems; see [Kiw92].

Let us now show how to employ Procedure 13.4 as a subroutine for implementing Step 4' of Algorithm 2.2; cf. §§10 and 11. Suppose Procedure 13.4 is called to find a point in $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\rm lev})$, starting from $\hat{x}^1 = x^k$ (with $\epsilon_{\rm tol} = 0$). Then it may be exited at any iteration $\bar{n} \geq 1$ also at Step 6. Specifically, in view of §4, we may take $\tilde{D} = r_k = (\bar{D}^2 - \rho_k)^{1/2}$, and Step 5 may use the additional test $(\tilde{D} - |\tilde{x}^{n+1} - x^{k(l)+1}|)^2 > r_k^2 - \tilde{\rho}_{n+1}$. Upon termination at Step j say, set $z^k = \hat{x}^{\bar{n}}$, $\rho^k_\phi = \hat{\rho}_{\bar{n}}$ if $j = 2$, $\rho^k_\phi = \infty$ if $j = 4$ or 5, and $z^k = \tilde{x}^{\bar{n}+1}$ and $\rho^k_\phi = \tilde{\rho}_{\bar{n}+1}$ if $j = 6$. Then z^k and ρ^k_ϕ satisfy (10.1) (cf. §4). The easiest way to ensure that $\rho^k_\phi \geq t_{\min}(2 - t_{\max})d^2_{\mathcal{L}(f^k, f^k_{\rm lev})}(x^k)$ consists in taking $i_1 \in \operatorname{Arg\,max}_i s^1_i/|a^i|$ at Step 1, since then $|\tilde{y}^1 - \tilde{x}^1| = d_{\mathcal{P}_{J^1}}(\tilde{x}^1) \geq d_{\mathcal{L}(f^k, f^k_{\rm lev})}(x^k)$. (In fact the usual choice yields $|\tilde{y}^1 - \tilde{x}^1| \geq (f(x^k) - f^k_{\rm lev})/\max_{j\in J^k}|g^j|$, and this suffices; cf. Lemma 3.1.) Thus, if desired, *only one* iteration of Procedure 13.4 may be executed, but more iterations will yield better z^k and ρ^k_ϕ for Algorithm 2.2. Note that *if Step 6 always exits* then we may set $n = k$ and $\tilde{x}^n = x^k$ at Step 0, terminating with $x^{k+1} = \hat{x}^{k+1}$ at Steps 4 or 5, or $z^k = \tilde{x}^{k+1}$ and $y^k = \tilde{y}^k$ at Step 6.

Remark 13.9. As in §12, the final $\lambda^{\bar{n}}$ may be used for subgradient selection or aggregation. Note that selective aggregation corresponds to dropping from A_{I^n} one column aggregated into another, thus retaining the crucial property $\langle a^i, a^j\rangle \leq 0\ \forall i \neq j$ in the new $I^{\bar{n}}$ (in contrast to total aggregation that replaces one column by a convex combination of all columns). Of course, the final $I^{\bar{n}}$ may become the initial I^0 on the next call to Procedure 13.4, and the final matrix factorization should be used in a hot start, e.g., if only $f^k_{\rm lev}$ has changed. We may add that most matrix factorizations can be updated to reflect selective aggregation.

Remark 13.10. The following modification is useful when $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\rm lev})$. Unless $f^k_{\rm lev} \equiv f^*$ is employed, it suffices to discover that $f^k_{\rm lev} \leq f^*$, since then $f^{k+1}_{\rm low} = f^k_{\rm lev}$ can be used without impairing the preceding efficiency results. To this end, Step 3 may choose *any* I^n such that $i_n \in I^n$, $s^n_i \geq 0$ and $\langle a^i, a^j\rangle \leq 0\ \forall i \neq j$, $i, j \in I^n$. Indeed, suppose $f^k_{\rm lev} > f^*$. By (2.12), $\hat{f}^k(x^*) \leq f^*$ for any $x^* \in S^*$ and, since $\hat{f}^k = \max_{j\in J^k} f^j$ and $\langle a^i, \cdot\rangle - b_i = f^j(\cdot) - f^k_{\rm lev}$ for suitable i and j, we have $\langle a^i, x^*\rangle - b_i \leq \hat{f}^k(x^*) - f^k_{\rm lev} < 0 \leq s^n_i = \langle a^i, \tilde{x}^n\rangle - b_i\ \forall i \in I^n$, so $A^T_{I^n}(x^* - \tilde{x}^n) < 0$ and $\operatorname{rank} A_{I^n} = |I^n|$ by Lemma 13.3(iii). Therefore, if $\operatorname{rank} A_{I^n} < |I^n|$ is revealed by any factorization then $f^k_{\rm lev} \leq f^*$ and Algorithm 2.2 may set $f^{k+1}_{\rm low} = f^k_{\rm lev}$.

Extending [Shc92], we now describe an *orthogonal surrogate projection* (OSP) version of Procedure 13.4 that sets $\tilde{y}^n = P_{\tilde{\mathcal{P}}_n}(\tilde{x}^n)$ for $\tilde{\mathcal{P}}_n = \{x : \langle q^j, x\rangle \le c_j, j \in \tilde{J}^n\}$, where each inequality is a surrogate of $A^T x \le b$ (so that $\mathcal{P} \subset \tilde{\mathcal{P}}_n$), the system $\mathcal{Q}_{\tilde{J}^n} = \{q^j\}_{j\in\tilde{J}^n}$ is orthonormal and $\tilde{J}^n \subset \{1{:}n\}$. Here $A^T_{\tilde{J}^{n-1}} x \le b_{\tilde{J}^{n-1}}$ is replaced by the *accumulated* surrogates $\langle q^j, x\rangle \le c_j$, $j \in \tilde{J}^{n-1}$, at Step 3 in constructing the new surrogate $\langle q^n, x\rangle \le c_n$ via orthogonalization as in Lemma 13.2. Specifically, at Step 0 set $\tilde{J}^0 = \emptyset$. At Step 3 set $\tilde{J}^{n-1} = \{j \in \tilde{J}^{n-1} : \langle a^{i_n}, q^j\rangle \le 0\}$ and $\tilde{J}^n = \tilde{J}^{n-1} \cup \{n\}$. At Step 4 set

$$\tilde{q}^n = a^{i_n} - \sum_{j\in\tilde{J}^{n-1}} \left\langle a^{i_n}, q^j \right\rangle q^j = 0 \quad \text{and} \quad q^n = \tilde{q}^n/|\tilde{q}^n| \quad \text{if} \quad \tilde{q}^n \neq 0, \tag{13.6a}$$

print "$\mathcal{P} = \emptyset$" and terminate if $\tilde{q}^n = 0$; otherwise set $\sigma^n_{\tilde{J}^{n-1}} = (1 - \tilde{t}_{n-1})\sigma^{n-1}_{\tilde{J}^{n-1}}$,

$$(c_n, \sigma^n_n) = [(b_{i_n}, s^n_{i_n}) - \sum_{j\in\tilde{J}^{n-1}} \left\langle a^{i_n}, q^j \right\rangle (c_j, \sigma^n_j)]/|\tilde{q}^n|, \tag{13.6b}$$

$\tilde{d}^n = -Q_{\tilde{J}^n}\sigma^n_{\tilde{J}^n}$ and $\tilde{y}^n = \tilde{x}^n + \tilde{d}^n$, and choose $\tilde{t}_n \le 1$. Here $Q_{\tilde{J}^n} = [Q_{\tilde{J}^{n-1}}, q^n]$, where $Q_{\tilde{J}^{n-1}}$ is the $N \times |\tilde{J}^{n-1}|$ orthonormal matrix corresponding to $\mathcal{Q}_{\tilde{J}^{n-1}}$.

To validate this modification, suppose $\mathcal{P} \neq \emptyset$, $Q_{\tilde{J}^{n-1}}$ is orthonormal, $\sigma^n_{\tilde{J}^{n-1}} = Q^T_{\tilde{J}^{n-1}}\tilde{x}^n - c_{\tilde{J}^{n-1}} \ge 0$ and $Q^T_{\tilde{J}^{n-1}} x \le c_{\tilde{J}^{n-1}}$ is a surrogate of $A^T x \le b$, so that $\mathcal{P} \subset \mathcal{P}_n = \tilde{x}^n + \mathcal{D}_n$, where $\mathcal{D}_n = \{d : Q^T_{\tilde{J}^{n-1}} d \le -\sigma^n_{\tilde{J}^{n-1}}, \langle a^{i_n}, d\rangle \le -s^n_{i_n}\}$. Also $s^n_{i_n} > 0$ and $Q^T_{\tilde{J}^{n-1}} a^{i_n} \le 0$ by construction. Hence, replacing $\mathcal{D}_I$ by $\mathcal{D}_n \neq \emptyset$ in Lemma 13.3, we deduce that $\operatorname{rank}[Q_{\tilde{J}^{n-1}}, a^{i_n}] = \hat{m} := |\tilde{J}^n|$, so $\tilde{q}^n \neq 0$ and by construction $[Q_{\tilde{J}^{n-1}}, a^{i_n}] = Q_{\tilde{J}^n} R$, where $Q_{\tilde{J}^n}$ is orthonormal, $r_{jj} = 1$ and $r_{j\hat{m}} = \langle a^{i_n}, q^j\rangle \le 0$, $j = 1{:}\hat{m}-1$, $r_{\hat{m}\hat{m}} = |\tilde{q}^n|$, and the remaining $r_{ij} = 0$. Then by (13.6), $\langle q^n, x\rangle \le c_n$ is a surrogate of $Q^T_{\tilde{J}^{n-1}} x \le c_{\tilde{J}^{n-1}}$, $\langle a^{i_n}, x\rangle \le b_{i_n}$ (and hence of $A^T x \le b$) with $\sigma^n_n = \langle q^n, \hat{x}^n\rangle - c_n > 0$, and $R^{-T}[(\sigma^n_{\tilde{J}^{n-1}})^T, s^n_{i_n}]^T = \sigma^n_{\tilde{J}^n}$. Therefore, since $\tilde{d}^n = -Q_{\tilde{J}^n}\sigma^n_{\tilde{J}^n}$, we deduce from Lemma 13.3(v) applied to $\mathcal{D}_n$ that $Q^T_{\tilde{J}^n}\tilde{d}^n = -\sigma^n_{\tilde{J}^n}$,

$$\mathcal{D}_n \subset \tilde{\mathcal{D}}_n := \{\, d : Q^T_{\tilde{J}^n} d \le -\sigma^n_{\tilde{J}^n} \,\} \quad \text{and} \quad \tilde{d}^n = P_{\mathcal{D}_n}(\tilde{x}^n) = P_{\tilde{\mathcal{D}}_n}(\tilde{x}^n). \tag{13.7}$$

Thus $\tilde{y}^n = P_{\mathcal{P}_n}(\tilde{x}^n) = P_{\tilde{\mathcal{P}}_n}(\tilde{x}^n)$, where $\mathcal{P}_n \subset \{x : \langle a^{i_n}, x\rangle \le b_{i_n}\}$ and $\tilde{\mathcal{P}}_n = \tilde{x}^n + \tilde{\mathcal{D}}_n$. Using $Q^T_{\tilde{J}^n}\tilde{d}^n = -\sigma^n_{\tilde{J}^n}$, $\tilde{x}^{n+1} = \tilde{x}^n + \tilde{t}_n\tilde{d}^n$ and $\tilde{t}_n \le 1$ gives $\sigma^{n+1}_{\tilde{J}^n} = Q^T_{\tilde{J}^n}\tilde{x}^{n+1} - c_{\tilde{J}^n} = \sigma^n_{\tilde{J}^n} + \tilde{t}_n Q^T_{\tilde{J}^n}\tilde{d}^n = (1 - \tilde{t}_n)\sigma^n_{\tilde{J}^n} \ge 0$. Also $Q^T_{\tilde{J}^n} x \le c_{\tilde{J}^n}$ is a surrogate of $A^T x \le b$. Hence one may use induction to show that this OSP version shares the convergence properties of the original one.

Note that if $\tilde{t}_{n-1} = 1$ then $\sigma^n_{\tilde{J}^{n-1}} = 0$ and $\tilde{d}^n = -\sigma^n_n q^n = -s^n_{i_n} q^n/|\tilde{q}^n|$ as in Remark 13.8. Thus $\tilde{y}^n$ is the projection of $\tilde{x}^n$ on the 'orthogonalized' surrogate $\langle q^n, x\rangle \le c_n$. Also the preceding validation would go through if, to save work, we increased $c_{\tilde{J}^{n-1}}$ by $\sigma^n_{\tilde{J}^{n-1}} \neq 0$, i.e., enlarged $\mathcal{P}_n$ to get $\sigma^n_{\tilde{J}^{n-1}} = 0$ and $\tilde{d}^n = -s^n_{i_n} q^n/|\tilde{q}^n|$ as in Remark 13.7.

Step 3 could construct more than one surrogate. Specifically, if $s^n_i > 0$ and $Q^T_{\tilde{J}^n} a^i \le 0$ for some i then Step 3 could append to $Q^T_{\tilde{J}^n} x \le c_{\tilde{J}^n}$ another surrogate derived from $\langle a^i, x\rangle \le b_i$ and $Q^T_{\tilde{J}^n} x \le c_{\tilde{J}^n}$ as in (13.6), and this may be repeated for other violated constraints. Again, Lemma 13.3 validates this extension.

To improve accuracy, *iterative refinement* of the form $\tilde{y}^n \leftarrow \tilde{y}^n + Q_{\tilde{J}^n}(c_{\tilde{J}^n} - Q^T_{\tilde{J}^n}\tilde{y}^n)$ or $\tilde{d}^n \leftarrow \tilde{d}^n + Q_{\tilde{J}^n}(\sigma^n_{\tilde{J}^n} - Q^T_{\tilde{J}^n}\tilde{d}^n)$ may be used at some iterations, and q^n should be *reorthogonalized* with respect to $Q_{\tilde{J}^{n-1}}$; see, e.g., [Bjö90, DGKS76].

When the OSP procedure is called repeatedly with $\mathcal{P} = \mathcal{L}(\hat{f}^k, f^k_{\rm lev})$, we may generate a vector $\check{c}_{j^n} > 0$ recursively via $\check{c}_n = (1 - \sum_{j\in \check{J}^{n-1}} \langle a^{i_n}, q^j\rangle \check{c}_j)/|\tilde{q}^n|$. Then, by induction on the surrogate construction (13.6), there exist $\nu_j \geq 0$ such that

$$(q^n, c_n) = \sum_{j=1}^{n} \nu_j (a^{i_j}, b_{i_j}) \quad \text{and} \quad \check{c}_n = \sum_{j=1}^{n} \nu_j \geq \nu_n = 1/|\tilde{q}^n| > 0, \tag{13.8}$$

so dividing by $\check{c}_n$ yields $(\langle q^n, \cdot\rangle - c_n)/\check{c}_n \in \mathrm{co}\{\langle a^i, \cdot\rangle - b_i\}_{i=1}^m \in \mathrm{co}\{f^j\}_{j\in J^k} - f^k_{\rm lev}$, and hence $\mathrm{co}\{(q^j, c_j)/\check{c}_j\}_{j\in \check{J}^n} \in \mathrm{co}\{(a^i, b_i)\}_{i=1}^m$. Thus, when $f^k_{\rm lev}$ changes to $f^{k+1}_{\rm lev}$ on the next call then $(f^{k+1}_{\rm lev} - f^k_{\rm lev})\check{c}_{j^n}$ should be added to c_{j^n} and $\sigma^n_{j^n}$. If $x^{k+1} \neq \tilde{x}^n$ (e.g., due to $x^{k+1} = P_S(\tilde{x}^n)$), then $\sigma^n_{j^n} = Q^T_{j^n}\tilde{x}^1 - c_{j^n}$ must be recomputed for $\tilde{x}^1 = x^{k+1}$. Alternatively, using $Q^T_{j^n}x - c_{j^n} = Q^T_{j^n}(x - \tilde{x}^n) + \sigma_{j^n}$, we may update $\sigma^n_{j^n} \leftarrow \sigma^n_{j^n} + Q^T_{j^n}(x^{k+1} - \tilde{x}^n)$ (then c_{j^n} is not required). Next, dropping j with $\sigma^n_j < 0$ from $\check{J}^n$, if any, a hot start can proceed as if $\check{J}^0$ were $\check{J}^n$. Note that Remark 13.10 holds for the OSP version with A_{I^n} replaced by $[Q_{j^{n-1}}, a^{i_n}]$ (since $A^T x^* < b \Rightarrow Q^T_{j^{n-1}}x^* - c_{j^{n-1}} < 0 \leq \sigma^n_{j^{n-1}} = Q^T_{j^{n-1}}\tilde{x}^n - c_{j^{n-1}}$ by (13.8)).

Remark 13.11. By the preceding argument, the *surrogate linearizations*

$$\tilde{f}^j(\cdot) = (\langle q^j, \cdot\rangle - c_j)/\check{c}_j + f^k_{\rm lev} = (\langle q^j, \cdot - \tilde{x}^n\rangle + \sigma^n_j)/\check{c}_j + f^k_{\rm lev}, \quad j \in \check{J}^n, \tag{13.9}$$

satisfy $\tilde{f}^j \in \mathrm{co}\{f^j\}_{j\in J^k} \subset \Phi$. Hence they may be used as any other linearizations of f. For instance, no additional storage is required if, at Step 6, $\tilde{f}^n$ replaces the f^j corresponding to $\langle a^{i_n}, \cdot\rangle - b_{i_n}$. Also $\lambda^n_j = \check{c}_j\sigma^n_j$ are Lagrange multipliers for $\tilde{f}^j$, $j \in \check{J}^n$, since $\tilde{d}^n = P_{\check{\mathcal{D}}_n}(\tilde{x}^n)$ with $\check{\mathcal{D}}_n = \{d : \langle q^j/\check{c}_j, d\rangle \leq -\sigma^n_j/\check{c}_j, j \in \check{J}^n\}$, whereas σ^n_j are Lagrange multipliers for $\tilde{d}^n = P_{\tilde{\mathcal{D}}_n}(\tilde{x}^n)$ in (13.7). Thus $\lambda^n_{j^n}$ can be used for selective aggregation as in Remark 13.9, and normalization of the aggregated column ensures orthonormality of the new Q_{j^n}.

Remark 13.12. Consider the simplest case where Step 6 always exits, so that we may let $n = k$ and $\tilde{x}^n = x^k$. Suppose $\mathcal{P} = \mathcal{L}(\phi^k, f^k_{\rm lev})$ with $\phi = \max\{f^k, \psi^{k-1}\}$ as for (12.2), where ψ^{k-1} is the previous aggregate satisfying $\psi^{k-1}(\tilde{x}^n) = f^k_{\rm lev}$, so that a hot start occurs from $\check{J}^{n-1} = \{n-1\}$ using $\tilde{q}^{n-1} = g^{k-1}_\psi$, $q^{n-1} = \tilde{q}^{n-1}/|\tilde{q}^{n-1}|$, $c_{n-1} = 1/|\tilde{q}^{n-1}|$ and $\sigma^n_{n-1} = 0$. Assume $\langle g^k, g^{k-1}_\psi\rangle < 0$. By simple calculation, either infeasibility is detected at Steps 4 or 5, or Step 6 terminates with $\tilde{d}^n = -(f(x^k) - f^k_{\rm lev})\tilde{q}^n/|\tilde{q}^n|^2$, where $\tilde{q}^n = g^k - \langle g^k, g^{k-1}_\psi\rangle g^{k-1}_\psi/|g^{k-1}_\psi|^2$ has $|\tilde{q}^n|^2 = |g^k|^2 - \langle g^k, g^{k-1}_\psi\rangle^2/|g^{k-1}_\psi|^2$ (use $\tilde{d}^n = -\sigma^n_n q^n$, $\sigma^n_n = s^n_{i_n}/|\tilde{q}^n|$ and $s^n_{i_n} = f(x^k) - f^k_{\rm lev}$). Also, since $\sigma^n_{n-1} = 0$, the aggregate ψ^k of f^k and ψ^{k-1} coincides with $\tilde{f}^n$. Note that for $\langle g^k, g^{k-1}_\psi\rangle \geq 0$ we would get $\tilde{q}^n = g^k$ (as if $\phi^k = f^k$), and that if we had $\psi^{k-1}(\tilde{x}^n) > f^k_{\rm lev}$ then the same formulae would hold if we set $c^n_{n-1} = 0$. It is not suprising that *the same* $\tilde{d}^n$ and ψ^k would be produced via the original version of Procedure 13.4 restarted from $A_{I^{n-1}} \cdot - b_{I^{n-1}} \equiv \psi^{k-1} - f^k_{\rm lev}$. (Use $\alpha^k_\psi = 0$ in (12.3)-(12.4) to get $\lambda^k_k = (f(x^k) - f^k_{\rm lev})/|\tilde{q}^n|^2$ and $d^k = -\lambda^k_k\tilde{q}^n = \tilde{d}^n$.) We add that $\psi^k(x^{k+1}) \geq f^{k+1}_{\rm lev}$ if $t_k \leq 1$ (cf. Lemma 14.1(v)).

We may add that the method of [Shc79, Shc92] corresponds to a version of Algorithm 2.2 that attempts to solve the inequality $f(x) \leq 0$. It sets $f^k_{\rm lev} = f^k_{\rm low} = 0$ and finds $x^{k+1} = \tilde{x}^2$

via one iteration of a simplified OSP version of Procedure 13.4 that starts with $\tilde{x}^1 = x^k$ and $f^k(x) \leq 0$ appended to the accumulated surrogates, and exits at Step 6, unless infeasibility is detected earlier, in which case it terminates. First, it sets $\tilde{t}_n = 1$ for all n, whereas our OSP version allows smaller stepsizes that may be useful at initial iterations. Second, it expresses $\tilde{d}^n$ as $\tilde{d}^n = -|\tilde{a}^n|^2 q^n / \langle \tilde{a}^n, q^n \rangle$, where $\tilde{a}^n = s^n_{i_n} a^{i_n} / |a^{i_n}|^2$, so $\tilde{d}^n = -s^n_{i_n} q^n / |\tilde{q}^n|$ from $\langle a^{i_n}, \tilde{q}^n \rangle = |\tilde{q}^n|^2$ by orthogonality. (In fact it replaces a^{i_n} by $\tilde{a}^n$ in calculating q^n, but this does not affect Q_{J^n}.) Third, it does not compute c_{J^n} and $\check{c}_{J^n}$, thus preventing iterative refinement and hot starts that would be necessary for handling the additional constraint $x \in S$ (except when S is a flat [Shc87]). Fourth, both methods should cope with the instability of the Gram-Schmidt process. Periodic resets to $\check{J}^n = \{n\}$ recommended in [Shc87, Shc92] slow down convergence. It seems better to employ iterative refinement and reorthogonalization in computing q^n. To sum up, our method appears competitive with that of [Shc92]. Finally, note that such methods project on sets $\mathcal{P}_n$ that may or may not be larger than $\mathcal{P}_{I^n}$ with $I^n = \{i_j\}_{j \in \check{J}^n}$. Thus it is not clear whether our OSP version could compete with, e.g., a Cholesky-based implementation of Procedure 13.4. We note that encouraging numerical results were obtained in [Ceg92] by a method that combines greatly simplified versions of Algorithm 2.2 and Procedure 13.4 for solving a consistent inequality $f(x) \leq 0$.

14 Conjugate subgradient techniques

In this section we use the dual framework of §9 for extending some conjugate subgradient (CS) techniques; see, e.g., [Brä91, CFM75, KiA90, SaKS7, SKRS7, ShU89, Sho79].

First, we identify surrogate linearizations of f that may be generated via CS methods.

Lemma 14.1. *Let $0 < \mu \leq 1$. Suppose iteration $k-1$ provides an affine model ψ^{k-1} of f_S of the form $\psi^{k-1}(\cdot) = \psi^{k-1}(x^k) + \left\langle g^{k-1}_\psi, \cdot - x^k \right\rangle$ with $\psi^{k-1}(x^k) \geq f^k_{\text{lev}}$ such that $\psi^{k-1} \in \Phi$ if $f^k_{\text{lev}} \geq f^*$. Let $\check{\psi}^{k-1} = \psi^{k-1} + f^k_{\text{lev}} - \psi^{k-1}(x^k)$ denote a shifted version of ψ^{k-1} such that $\check{\psi}^{k-1}(x^k) = f^k_{\text{lev}}$. The corresponding current models of f_S are given by $\hat{\phi}^k = \max\{f^k, \psi^{k-1}\}$ and $\check{\phi}^k = \max\{f^k, \check{\psi}^{k-1}\}$. Next, let $\check{\phi}^k = f(x^k) + \left\langle \check{g}^k_\phi, \cdot - x^k \right\rangle = f^k + \beta_k \left\langle g^{k-1}_\psi, \cdot - x^k \right\rangle$ be the current CS model of f_S, where $\check{g}^k_\phi = g^k + \beta_k g^{k-1}_\psi$ for $\beta_k \geq 0$ such that $|\check{g}^k_\phi| \leq |g^k|/\mu$, and let $\tilde{\phi}^k = (f^k + \beta_k \psi^{k-1})/(1+\beta_k)$ denote another CS-like model of f_S that is a convex combination of f^k and ψ^{k-1}. Finally, let $\check{\beta}_k = \arg\min_{\beta \geq 0} |g^k + \beta g^{k-1}_\psi|$. If $f^k_{\text{lev}} \geq f^*$ then:*

(i) *$\psi^{k-1}, \check{\psi}^{k-1} \in \Phi$, $\hat{\phi}^k, \check{\phi}^k \in \Phi^k_1$ and $\check{\phi}^k, \tilde{\phi}^k \in \Phi^k_\mu$ (cf. (9.2)). In particular,*

$$\begin{aligned} d_{\mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})}(x^k) &\geq d_{\mathcal{L}(\tilde{\phi}^k, f^k_{\text{lev}})}(x^k) = [f(x^k) - f^k_{\text{lev}} + \beta_k(\psi^{k-1}(x^k) - f^k_{\text{lev}})]/|\check{g}^k_\phi| \\ &\geq d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) = (f(x^k) - f^k_{\text{lev}})/|\check{g}^k_\phi| \geq \mu\kappa\Delta^k/L_f \end{aligned}$$

and $d_{\mathcal{L}(\hat{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$.

(ii) *$\check{\beta}_k = \left\langle g^k, -g^{k-1}_\psi \right\rangle_+ / |g^{k-1}_\psi|^2$. Moreover, $|\check{g}^k_\phi|^2 = |g^k|^2 - \left\langle g^k, g^{k-1}_\psi \right\rangle^2 / |g^{k-1}_\psi|^2$ if $\beta_k = \check{\beta}_k$ and $\left\langle g^k, g^{k-1}_\psi \right\rangle < 0$, $|\check{g}^k_\phi| \leq |g^k|$ if $\beta_k \leq 2\check{\beta}_k$, and $|\check{g}^k_\phi| \leq 2|g^k|$ if $\beta_k = |g^k|/|g^{k-1}_\psi|$.*

(iii) *$\beta_k \leq 2\check{\beta}_k \Rightarrow d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) \geq d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$; $\beta_k < 2\check{\beta}_k \Rightarrow d_{\mathcal{L}(\check{\phi}^k, f^k_{\text{lev}})}(x^k) > d_{\mathcal{L}(f^k, f^k_{\text{lev}})}(x^k)$.*

(iv) *If* $\left\langle g^k, g_\psi^{k-1}\right\rangle < 0$ *and* $\beta_k = \check\beta_k$ *then* $P_{\mathcal{L}(\check\phi^k, f_{\rm lev}^k)}(x^k) = P_{\mathcal{L}(\hat\phi^k, f_{\rm lev}^k)}(x^k)$.

(v) $\phi^k(x^{k+1}) \geq f_{\rm lev}^k$ *if* $t_k \leq 1$ *and* $\phi^k(x^k) > f_{\rm lev}^k$; *e.g.,* $\phi^k = \check\phi^k$, $\tilde\phi^k$, $\hat\phi^k$ *or* $\dot\phi^k$ *at Step* 4. *Conversely, if* $\phi^k = \check\phi^k$, $\tilde\phi^k$, $\hat\phi^k$ *or* $\dot\phi^k$ *and* $\mathcal{L}(\phi^k, f_{\rm lev}^k) = \emptyset$ *then* $f_{\rm lev}^k < f^*$.

Proof. (i) Let $x^* \in S^*$. Since $\psi^{k-1}(x^k) + \left\langle g_\psi^{k-1}, x^* - x^k\right\rangle = \psi^{k-1}(x^*) \leq f^*$ and $f^k(x^*) \leq f^*$, we have $\left\langle g_\psi^{k-1}, x^* - x^k\right\rangle \leq f^* - \psi^{k-1}(x^k) \leq 0$ and $\check\phi^k(x^*) = f^k(x^*) + \beta_k\left\langle g_\psi^{k-1}, x^* - x^k\right\rangle \leq f^*$. Next, $d_{\mathcal{L}(\check\phi^k, f_{\rm lev}^k)}(x^k) = (f(x^k) - f_{\rm lev}^k)/|\check g_\phi^k| \geq \kappa\Delta^k/(|g^k|/\mu) \geq \mu\kappa\Delta^k/L_f$, while $\tilde\phi^k(x) = \tilde\phi^k(x^k) + \left\langle \tilde g_\phi^k, x - x^k\right\rangle$ with $\tilde\phi^k(x^k) - f_{\rm lev}^k = [f^k(x^k) - f_{\rm lev}^k + \beta_k(\psi^{k-1}(x^k) - f_{\rm lev}^k)/(1+\beta_k)$ and $\tilde g_\phi^k = (g^k + \beta_k g_\psi^{k-1})/(1+\beta_k)$ yield $d_{\mathcal{L}(\tilde\phi^k, f_{\rm lev}^k)}(x^k) = [f(x^k) - f_{\rm lev}^k + \beta_k(f(x^k) - f_{\rm lev}^k)]/|\check g_\phi^k|$, so the conclusion follows from $\check\psi^{k-1} \leq \psi^{k-1}$, $f^k \leq \check\phi^k \leq \hat\phi^k$ and $\tilde\phi^k \leq \check\phi^k$.

(ii) Solve $\min_{\beta\geq 0}|g^k + \beta g_\psi^{k-1}|^2$, and use $|\check g^k| \leq |g^k| + \beta_k|g_\psi^{k-1}|$.

(iii) Invoke (i) with $|\check g^k| \leq |g^k|$ and $d_{\mathcal{L}(f^k, f_{\rm lev}^k)}(x^k) = (f(x^k) - f_{\rm lev}^k)/|g^k|$.

(iv) Use $|\check g_\phi^k|^2 = |g^k|^2 - \left\langle g^k, g_\psi^{k-1}\right\rangle^2/|g_\psi^{k-1}|^2$ and $\alpha_\psi^k = 0$ in (12.3)-(12.4) to get $\lambda_k^k = (f(x^k) - f_{\rm lev}^k)/|\check g_\phi^k|^2$ and $d^k = -\lambda_k^k \check g_\phi^k = P_{\mathcal{L}(\check\phi^k, f_{\rm lev}^k)}(x^k) - x^k$ with $J^k = \{k\}$ in (12.2).

(v) Using (2.8), $x^{k+1} = P_S(z^k)$, $x^k \in S$ and $t_k \leq 1$, we get $|x^{k+1} - x^k| \leq |z^k - x^k| = t_k|y^k - x^k| \leq |y^k - x^k|$. But $y^k = P_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k)$ with $\phi^k(x^k) > f_{\rm lev}^k$, so $f_{\rm lev}^k = \phi^k(y^k) \leq \phi^k(x^{k+1})$. □

Lemma 14.1 suggests the following *CS implementation* of Algorithm 2.2. Let $0 < \mu \leq 1$ and $\phi^1 = f^1$. At iteration $k \geq 2$, let $\psi^{k-1} = \phi^{k-1}$ and $\phi^k = \check\phi^k$ with $\beta_k \geq 0$ such that $|\check g_\phi^k| \leq |g^k|/\mu$ if $\psi^{k-1}(x^k) \geq f_{\rm lev}^k$, and $\beta_k = 0$ ($\phi^k = f^k$) otherwise. Then by induction, as in §9, we see that only the first terms of the constants in all the preceding efficiency estimates and the right side of (7.3) need be divided by μ^2, with $\mu = 1$ if $\beta_k \leq 2\check\beta_k$ $\forall k$; of course, $\Delta_{\rm lev}^k$ replaces $\kappa\Delta^k$ in Lemma 14.1 for the frozen level gaps of §7.

Note that by construction, $d^k = -(f(x^k) - f_{\rm lev}^k)\check g_\phi^k/|\check g_\phi^k|^2$ and if $d^{k-1} \neq 0$ then $d^{k-1}/|d^{k-1}| = -g_\psi^{k-1}/|g_\psi^{k-1}|$, so if $\beta_k = \check\beta_k$ and $\left\langle g^k, d^{k-1}\right\rangle > 0$ then $\check g_\phi^k = g^k - \left\langle g^k, d^{k-1}\right\rangle d^{k-1}/|d^{k-1}|^2$ and $\left\langle d^k, d^{k-1}\right\rangle = 0$. These CS relations correspond to those of the methods in [Brä91, CFM75, KiA90, ShUS9, Sho79], which set $f_{\rm lev}^k = f^*$ and $t_k \leq 1$. Incidentally, when $t_k \leq 1$ and $f_{\rm lev}^{k+1} \leq f_{\rm lev}^k$ then $\phi^k(x^{k+1}) \geq f_{\rm lev}^{k+1}$ by Lemma 14.1(v), so such methods can skip computing $\psi^{k-1}(x^k)$ ($\geq f^*$); this is the main reason for choosing $t_k \leq 1$. Usually $\beta_k \leq 2\check\beta_k$ is advocated; with the choice of $\beta_k = |g^k|/|g_\psi^{k-1}|$ from [ShUS9], the direction d^k simply bisects the angle between $-g^k$ and d^{k-1}, and, in this sense, is an *average direction*. Since $|y - y^k|^2 \leq |y - x^k|^2 - d^2_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k)$ if $y \in S^*$, $y^k = P_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k)$ and $f_{\rm lev}^k \geq f^*$, Lemma 14.1(iii) augments the usual angle-based motivation for using $\check\phi^k$ instead of f^k, while Lemma 14.1(iv) complements results in [KiA90]. In particular, the CS implementation with $\beta_k = \check\beta_k$ corresponds to the simplest OSP implementation of Remark 13.12 with ψ^{k-1} replaced by $\check\psi^{k-1}$ to zero its σ_{n-1}^n, and $\check g_\phi^k = \tilde q^n$ obtained by orthogonalizing g^k and g_ψ^{k-1}.

Lemma 14.1 says that we may easily improve classical CS techniques by taking $\phi^k = \check\phi^k$, $\tilde\phi^k$ or $\hat\phi^k$ instead of $\phi^k = \check\phi^k$ to increase $d_{\mathcal{L}(\phi^k, f_{\rm lev}^k)}(x^k)$; cf. Remark 3.9. In particular, $\tilde\phi^k$ is a convex combination of f^k and ψ^{k-1}, and other such combinations could be developed as in §9. It seems, however, that $\hat\phi^k = \max\{f^k, \psi^{k-1}\}$ is preferable anyway. First, Lemmas

9.4 and 14.1 show that $\hat{\phi}^k$ is best in terms of efficiency estimates. Second, the resulting choice of $\psi^{k-1} = \phi^{k-1}$ and $\hat{\phi}^k = \max\{f^k, \phi^{k-1}\}$ corresponds to the aggregate subgradient implementation of §12, which, in contrast to the other CS choices, does not require $t_k \leq 1$ and does not need to resort to the poorest model $\phi^k = f^k$ when $\phi^{k-1}(x^k) < f^k_{\rm lev}$ or $\left\langle g^k, g^{k-1}_\psi \right\rangle \geq 0$. Third, it involves little additional work (cf. (12.3)). Last, but not least, it is simpler conceptually. Incidentally, the choices of $\phi^k = \check{\phi}^k$ and $\phi^k = \hat{\phi}^k$ may be compared in dual terms by noting that $\check{\beta}_k = \arg\max_{\beta \geq 0}(f(x^k) - f^k_{\rm lev})/|g^k + \beta g^{k-1}_\psi|$ and $(\lambda^k_k, \lambda^k_\psi) \in \operatorname{Arg\,max}\{[\lambda_k(f(x^k) - f^k_{\rm lev}) + \lambda_\psi(\psi^{k-1}(x^k) - f^k_{\rm lev})]/|\lambda_k g^k + \lambda_\psi g^{k-1}_\psi| : \lambda_k \geq 0, \lambda_\psi \geq 0, \lambda_k + \lambda_\psi = 1\}$, where the second maximum $(= d_{\mathcal{L}(\hat{\phi}^k, f^k_{\rm lev})}(x^k))$ can be much greater.

An obvious extension of the CS techniques is to take $\phi^k = \max\{\hat{f}^k, \check{\phi}^k\}$ or $\max\{\hat{f}^k, \tilde{\phi}^k\}$ to increase $d_{\mathcal{L}(\phi^k, f^k_{\rm lev})}(x^k)$. Further, more than one CS steps can be made as in Remark 12.2.

15 Constraint modelling

Since Algorithm 2.2 minimizes f on S, ϕ^k should be chosen to model the extended objective $f_S = f + \delta_S$ and not just f alone. Failure to do so may result in severe deficiencies, as shown in the following simple example.

Example 15.1. Let $N = 2$, $S = [0,1] \times [0,1]$ and $f(x) = \epsilon x_1 + x_2$, where $0 < \epsilon < 1$ is a small parameter. Then $S^* = \{(0,0)\}$, $f^* = 0$, $\operatorname{diam}(S) = \sqrt{2}$ and $L_f = \sqrt{1+\epsilon^2}$. Let $x^1 = (1,0)$ and $t_{\min} = t_{\max} = 1$. The following facts are easy to verify by induction. The SPA (1.2) generates $x^k = ((1+\epsilon^2)^{1-k}, 0)$ and $f(x^k) = \epsilon(1+\epsilon^2)^{1-k}$ $\forall k$, i.e., its convergence is linear but very slow for small ϵ. The situation is even slightly worse for the SPLA (2.2) with $f^1_{\rm low} = f^*$, which yields $x^k = ([1 - \kappa\epsilon^2]/(1+\epsilon^2)^{k-1}, 0)$ $\forall k$. In contrast, Algorithm 2.2 with $f^1_{\rm low} = f^*$, $\bar{D} \geq \sqrt{2}$ and $\phi^k \equiv f^k + \delta_S$ gives $x^k = ((1-\kappa)^{k-1}, 0)$ $\forall k$, i.e., it is much faster for typical κ, independently of ϵ. In fact for $\kappa = 1$ (cf. Theorem 6.1) it terminates with $x^2 \in S^*$, being equivalent to the iteration (1.5).

Remark 3.9 and Example 15.1 suggest that the following modification of (2.2)

$$x^{k+1} = \arg\min\{\, |x - x^k|^2/2 : f(x^k) + \left\langle g_f(x^k), x - x^k \right\rangle \leq f^k_{\rm lev}, x \in S \,\} \qquad (15.1)$$

should be more efficient in practice. Supposing S is a box of the form $[x^{\rm low}, x^{\rm up}]$, let $x(\nu) = \arg\min_{x \in S}\{|x - x^k|^2/2 + \nu \left\langle g^k, x \right\rangle\}$ $\forall \nu \geq 0$. Then $x^{k+1} = x(\hat{\nu})$, where $\hat{\nu} \geq 0$ solves the equation $h(\nu) \equiv f(x^k) + \left\langle g^k, x(\nu) - x^k \right\rangle - f^k_{\rm lev} = 0$. Since $x_i(\nu) = \max\{\min[x^{\rm low}_i, x^k_i - \nu g^k_i], x^{\rm up}_i\}$, $i = 1{:}N$, and h is nonincreasing and piecewise linear, $\hat{\nu}$ is easy to compute.

Of course, projecting on S is easy only if S is simple enough, e.g., a Cartesian product of boxes, simplices, balls, ellipsoids, cylinders, etc. Additional linear constraints may complicate the projections; e.g., for $\phi^k = \hat{f}^k + \delta_S$ we must find

$$y^k = \arg\min\{\, |x - x^k|^2/2 : f^j(x) \leq f^k_{\rm lev}, j \in J^k, x \in S \,\}. \qquad (15.2)$$

Fortunately, accurate projections are not really necessary. For instance, (15.2) can be implemented approximately by projecting *cyclically* on $\mathcal{L}(\hat{f}^k, f^k_{\rm lev})$ and S as in Remark 12.2,

possibly with inexact projections on $\mathcal{L}(\hat{f}^k, f^k_{\rm lev})$ being performed via the methods of §§12, 13 and 14. Also if S is polyhedral then (15.2) is just (12.1) augmented with the inequalities of S, so it can be solved approximately by several steps of these methods.

It is crucial to observe that even if S is not polyhedral, it may still be *linearized* via inequalities generated in the course of calculations. First, such inequalities may be recovered *geometrically* by noting that $S \subset \mathcal{H}^k := \{x : \langle z^{k-1} - x^k, x - x^k \rangle \leq 0\}$ from $x^k = P_S(z^{k-1})$. Hence we may replace S in (15.2) by $S^k = \cap_{j \in J^k_S} \mathcal{H}^j$ with $J^k_S \subset \{2\!:\!k\}$. Second, similar inequalities may be generated *analytically* if $S = \{x : F(x) \leq 0\}$ say, where $F : \mathbb{R}^N \to \mathbb{R}$ is convex, and we can find its linearization $\bar{F}(\cdot; x) = F(x) + \langle g_F(x), \cdot - x \rangle$ with $g_F(x) \in \partial F(x)$ for any x. Then we may use $\mathcal{H}^k = \{x : \bar{F}(x; x^k) \leq 0\}$ as above. In other words, we may accumulate $\mathcal{H}^j$ for the model δ_{S^k} of δ_S in the same way as we use f^j in the model $\hat{f}^k$ of f, so that $f + \delta_S$ is approximated by $\phi^k = \hat{f}^k + \delta_{S^k}$, with $\delta_{S_k} \leq \delta_S$ from $S \subset S^k$. To save storage, some of the inequalities defining S^k may be aggregated as in §12.

The following observation is useful for the dual models of §9. If we take $\hat{\phi}^k = \hat{f}^k + \delta_S$ for a polyhedral $S = \{x : \langle a^i, x \rangle \leq b_i, i = 1\!:\!m\}$ say, then $\partial_{\epsilon_k} \hat{\phi}^k(x^k) = \{\partial_{\epsilon'} \hat{f}^k(x^k) + \partial_{\epsilon''} \delta_S(x^k) : \epsilon' + \epsilon'' \leq \epsilon_k, \epsilon', \epsilon'' \geq 0\}$ with

$$\partial_\epsilon \delta_S(x^k) = \{ \sum_{i=1}^m \nu_i a^i : \sum_{i=1}^m \nu_i (b_i - \langle a^i, x^k \rangle) \leq \epsilon, \nu_i \geq 0, i = 1\!:\!m \},$$

so $p^k = \arg\min\{|p|^2/2 : p \in \partial_{\epsilon_k} \hat{\phi}^k(x^k)\}$ of (9.5) can again be found via QP using (9.1). Next, letting $I^k = \{i : \langle a^i, x^k \rangle = b_i\}$ and $\hat{S}^k = \{x : \langle a^i, x \rangle \leq b_i, i \in I^k\}$, consider the simpler model $\hat{\phi}^k = f^k + \delta_{\hat{S}^k}$. Clearly, $\hat{\phi}^k \leq f_S$, $\hat{\phi}^k(x^k) = f(x^k)$ and $\partial_{\epsilon_k} \hat{\phi}^k(x^k) = g^k + \partial \delta_{\hat{S}^k}(x^k)$ for *any* $\epsilon_k \geq 0$, i.e., p^k does not depend on ϵ_k. Hence by Lemma 9.4, $\epsilon_k = 0$ gives the 'optimal' dual method with $d^k = y^k - x^k = -(f(x^k) - f^k_{\rm lev}) p^k / |p^k|^2 = P_{\mathcal{L}(\hat{\phi}^k, f^k_{\rm lev})}(x^k) - x^k$ if $p^k \neq 0$; otherwise $f^k_{\rm lev} < f^*$ by Lemma 9.2. (Thus it is not suprising that this dual method behaves like the primal one in Example 15.1.) Additional insight may be gained as follows. The cones $C = \text{cone}\{a^i\}_{i \in I^k}$ and $C^- = \{x : \langle a^i, x \rangle \leq 0, i \in I^k\}$ provide the classical orthogonal decomposition

$$g = P_C(g) + P_{C^-}(g), \quad P_C(g) \perp P_{C^-}(g), \quad \forall g \in \mathbb{R}^N,$$

since $P_{C^-}(g) = \arg\min\{|x - g|^2/2 : A^T_{I^k} x \leq 0\}$ has multipliers $\lambda_{I^k} \in \text{Arg}\min\{|A_{I^k} \lambda_{I^k} - g|^2/2 : \lambda_{I^k} \geq 0\}$ satisfying $P_C(g) = A_{I^k} \lambda_{I^k}$ and $\lambda^T_{I^k} A^T_{I^k} P_{C^-}(g) = 0$ by the K-T conditions, so for $g = -g^k$,

$$-p^k = -g^k - P_{\mathcal{N}_S(x^k)}(-g^k) = P_{\mathcal{T}_S(x^k)}(-g^k), \tag{15.3}$$

where $\mathcal{N}_S(x^k) = \partial \delta_S(x^k) = C$ and $\mathcal{T}_S(x^k) = C^-$ are the normal and tangent cones to S at x^k respectively. Hence $-p^k$ and d^k are feasible directions for S at x^k. Thus if t_k is sufficiently small then $z^k = x^k + t_k d^k \in S$, so that one may take $x^{k+1} = z^k$, skipping its projection on S. This motivates a similar technique in [KiU89] (with $f^k_{\rm lev} = f^*$), but small stepsizes may yield slow convergence.

Of course, if S is not polyhedral, then the preceding construction may employ its accumulated approximation S^k. A simple but useful example is given in the following

Lemma 15.2. *Suppose* $\mathcal{H}^k = \{x : \langle a^k, x - x^k \rangle \le 0\}$ *is a nontrivial outer approximation to* S *at* x^k*; e.g.,* $a^k = z^{k-1} - x^k \ne 0$*. Let* $\hat{\gamma}_k = \arg\min_{\gamma \ge 0} |g^k + \gamma a^k|$*, choose* $0 \le \gamma_k \le 2\hat{\gamma}_k$ *and let* $g_S^k = g^k + \gamma_k a^k$*. Then* $f_S^k(\cdot) = f^k(\cdot) + \gamma_k \langle a^k, \cdot - x^k \rangle = f(x^k) + \langle g_S^k, \cdot - x^k \rangle$ *is a valid linearization of* f_S*, i.e.,* $f_S^k \in \Phi_1^k$*, satisfying* $f_S^k(x^*) \le f^* \ \forall x^* \in S^*$*, and* $d_{\mathcal{L}(f_S^k, f_{\text{lev}}^k)}(x^k) \ge d_{\mathcal{L}(f^k, f_{\text{lev}}^k)}(x^k)$*, with strict inequality if* $0 < \gamma_k < 2\hat{\gamma}_k$*. In particular, if* $\langle g^k, a^k \rangle < 0$ *and* $\gamma_k = \hat{\gamma}_k = -\langle g^k, a^k \rangle / |a^k|^2$ *then* $g_S^k = g^k - \langle g^k, a^k \rangle a^k / |a^k|^2 = -P_{\mathcal{H}^k}(-g^k)$*.*

Proof. Clearly, $f_S^k(x^k) = f(x^k)$ and $f^k(x) \ge f_S^k(x) \ \forall x \in S \subset \mathcal{H}^k$, so $f_S^k(x^*) \le f^*$ if $x^* \in S^*$. As for the rest, solve $\min_{\gamma \ge 0} |g^k + \gamma a^k|^2$ and use $d_{\mathcal{L}(f_S^k, f_{\text{lev}}^k)}(x^k) = (f(x^k) - f_{\text{lev}}^k)/|g_S^k|$. □

In view of Lemma 15.2, we may replace g^k by $g_S^k \in \partial f_S(x^k)$, a conditional subgradient of f on S. In general, f_S^k is a worse model of f_S than $f^k + \delta_{\mathcal{H}^k}$, but it may be easier to handle.

We now extend the *average direction strategy* (ADS) of [ShUS9].

Lemma 15.3. *Suppose* $t_k \le 1$ *and* $\phi^k = \check{\phi}^k \in \Phi_\mu^k$*, where* $\check{\phi}^k = f(x^k) + \langle \check{g}_\phi^k, \cdot - x^k \rangle$ *is the CS model of Lemma* 14.1*. Let* $\psi^k = \check{\phi}^k(x^{k+1}) + \langle g_\psi^k, \cdot - x^{k+1} \rangle$*, where* $g_\psi^k = \check{g}_\phi^k + \gamma_k(z^k - x^{k+1})$ *for some* $\gamma_k \ge 0$*. Then* $\psi^k(x^{k+1}) \ge f_{\text{lev}}^k$ *and* $\psi^k \in \Phi$ *if* $f_{\text{lev}}^k \ge f^*$*. Moreover, if* $\gamma_k = \check{\gamma}_k := |\check{g}_\phi^k|^2 / (\check{\phi}^k(x^k) - f_{\text{lev}}^k) t_k$ *then* $\check{d}^k = -g_\psi^k / \gamma_k$*, where* $\check{d}^k = x^{k+1} - x^k$ *is the actual direction of motion which includes the effect of the projection operation. In particular, if* $\check{d}^k \ne 0$ *and* ψ^k *is used to define the next* $\phi^{k+1} = \check{\phi}^{k+1}$ *with* $\beta_{k+1} = |g^{k+1}|/|g_\psi^{k+1}|$ *then* $\check{g}_\phi^{k+1} = g^{k+1} - |g^{k+1}| \check{d}^k / |\check{d}^k|$*, i.e., the move* $y^{k+1} = x^{k+1} - (f(x^{k+1}) - f_{\text{lev}}^{k+1}) \check{g}_\phi^{k+1} / |\check{g}_\phi^{k+1}|^2$ *occurs along the average direction of* $-g^{k+1}$ *and* $\check{d}^k$*.*

Proof. If $f_{\text{lev}}^k \ge f^*$ and $x^* \in S^*$ then $\check{\phi}^k(x^{k+1}) + \langle \check{g}_\phi^k, x^* - x^{k+1} \rangle = \check{\phi}^k(x^*) \le f^*$ by (9.2) and $\langle z^k - x^{k+1}, x^* - x^{k+1} \rangle \le 0$ because $x^{k+1} = P_S(z^k)$ and $x^* \in S$, so $\psi^k(x^*) \le f^*$, while $\psi^k(x^{k+1}) = \phi^k(x^{k+1}) \ge f_{\text{lev}}^k$ by Lemma 14.1(v). Next, suppose $\gamma_k = \check{\gamma}_k$. Then $z^k - x^k = -t_k(\check{\phi}^k(x^k) - f_{\text{lev}}^k) \check{g}_\phi^k / |\check{g}_\phi^k|^2$ yields $-g_\psi^k / \gamma_k = z^k - x^k - (z^k - x^{k+1}) = \check{d}^k$. The rest follows by construction. □

Lemma 15.3 suggests the following *ADS version* of the CS implementation from §14. Let $0 < \mu \le 1$. At iteration k, let $\phi^1 = f^1$ if $k = 1$, otherwise use ψ^{k-1} to find $\phi^k = \check{\phi}^k$ with $\beta_k \ge 0$ such that $|\check{g}_\phi^k| \le |g^k|/\mu$ if $\psi^{k-1}(x^k) \ge f_{\text{lev}}^k$, and $\beta_k = 0$ otherwise; in both cases choose $\gamma_k \ge 0$ to construct ψ^k as in Lemma 15.3. In other words, instead of using $\psi^{k-1} = \phi^{k-1}$, the ADS version modifies ψ^{k-1} to include the effect of the projection operation. Clearly, the ADS version shares the efficiency estimates of the CS one's. On the other hand, the ADS version of a similar method in [ShUS9] (with $\gamma_k \equiv \check{\gamma}_k$) performed better than the standard CS version of [CFM75] (corresponding to $\beta_k = \check{\beta}_k$ in Lemma 14.1). (By the way, the ADS version was not validated theoretically in [ShUS9].) We note, however, that the arguments of §14 that favor $\hat{\phi}^k = \max\{f^k, \psi^{k-1}\}$ versus $\check{\phi}^k$ hold also for this modified form of ψ^{k-1}.

16 Conclusions

We have provided efficiency estimates for several new variants of subgradient relaxation methods for convex minimization. First, our theoretical framework extends and unifies the approaches of [KAC91, KuF90, LNN91] in the three schemes for underestimating the optimal value. Our fourth scheme with partial model minimization from §8 seems to be particularly promissing. Second, we have generalized the dual methods of [KuF90, LNN91], comparing them with primal ones and extending various acceleration techniques for subgradient methods, such as surrogate constraints, deepest surrogate cuts, simultaneous projections, orthogonal surrogate projections, conjugate subgradients and projected (conditional) subgradients. Finally, we have proposed to use subgradient aggregation and parallel projection methods for implementing our methods in the large-scale case.

Of course, some of our ideas have been inspired by other popular approaches [AHKS87, BaS81, HWC74, KKA87, SeS86, ShM88, ShU89], and may in turn be used to modify the methods given in these papers. For example, the concept of relying on subgradient aggregation to provide some 'conjugacy' (cf. §§12, 14 and 15) would enable the method of [ShU89] to use 'deeper cuts', thus enhancing faster convergence. We hope, therefore, that this paper will contribute to the development of other subgradient methods.

Acknowledgment. Most of this work was done during my six month stay at INRIA, Rocquencourt, in 1992, owing to the kind invitation by C. Lemarechal and the French Ministry for Research and Technology, whose financial support is gratefully acknowledged.

References

[Agm54] S. Agmon, *The relaxation method for linear inequalities*, Canad. J. Math. **6** (1954) 382–392.

[AhC89] R. Aharoni and Y. Censor, *Block-iterative projection methods for parallel computation of solutions to convex feasibility problems*, Linear Algebra Appl. **120** (1989) 165–175.

[AHKS87] E. Allen, R. Helgason, J. Kennington and B. Shetty, *A generalization of Polyak's convergence result for subgradient optimization*, Math. Programming **37** (1987) 309–317.

[BaG79] M. S. Bazaraa and J. J. Goode, *A survey of various tactics for generating Lagrangian multipliers in the context of Lagrangian duality*, European J. Oper. Res. **3**, no. 3 (1979) 322–338.

[BaS81] M. S. Bazaraa and H. D. Sherali, *On the choice of step size in subgradient optimization*, European J. Oper. Res. **7** (1981) 380–388.

[BGT81] R. B. Bland, D. Goldfarb and M. J. Todd, *The ellipsoid method: a survey*, Oper. Res. **29** (1981) 1039–1091.

[Bjö90] Å. Björck, *Least squares methods*, in Finite Difference Methods (Part 1) — Solution of Equations in R^n (Part 1), P. G. Ciarlet and J. L. Lions, eds., Handbook of Numerical Analysis, vol. I, North-Holland, Amsterdam, 1990, pp. 465–652.

[Brä91] U. Brännlund, *A generalized subgradient method based on the Polyak step*, in 2nd Stockholm Optimization Days, August 12–13, 1991, P. O. Lindberg, ed., Department of Mathematics, Royal Institute of Technology, Stockholm, 1991.

[Ceg92] A. Cegielski, *On a relaxation algorithm in convex optimization problems*, Tech. report, Institute of Mathematics, Higher College of Engineering, Zielona Góra, Poland, 1992.

[CFM75] P. M. Camerini, L. Fratta and F. Maffioli, *On improving relaxation methods by modified gradient techniques*, Math. Programming Stud. **3** (1975) 26–34.

[DGKS76] J. W. Daniel, W. B. Gragg, L. C. Kaufman and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt* qr *factorization*, Math. Comp. **30** (1976) 772–795.

[Dre83] Z. Drezner, *The nested ball principle for the relaxation method*, Oper. Res. **31** (1983) 587–590.

[Fle87] R. Fletcher, *Practical Methods of Optimization*, second ed., Wiley, Chichester, 1987.

[GMW91] P. E. Gill, W. Murray and M. H. Wright, *Numerical Linear Algebra and Optimization*, Addison-Wesley, Redwood City, CA, 1991.

[Gof78] J.-L. Goffin, *Nondifferentiable optimization and the relaxation method*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978, pp. 31–49.

[Gof81] ——, *Convergence results in a class of variable metric subgradient methods*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 283–326.

[GoT82] D. Goldfarb and M. J. Todd, *Modifications and implementation of the ellipsoid algorithm for linear programming*, Math. Programming **23** (1982) 1–19.

[GPR67] L. G. Gurin, B. T. Polyak and E. V. Raik, *The method of projections for finding a common point of convex sets*, Zh. Vychisl. Mat. i Mat. Fiz. **7**, no. 6 (1967) 1211–1228 (Russian). English transl. in U.S.S.R. Comput. Math. and Math. Phys. 7 (1967) 1–24.

[GVL89] G. H. Golub and C. F. Van Loan, *Matrix Computations*, second ed., The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[HWC74] M. Held, P. Wolfe and H. P. Crowder, *Validation of subgradient optimization*, Math. Programming **6**, no. 1 (1974) 62–88.

[IDP91] A. N. Iusem and A. R. De Pierro, *On the convergence of Han's method for convex programming with quadratic objective*, Math. Programming **52** (1991) 265–284.

[KAC91] S. Kim, H. Ahn and S.-C. Cho, *Variable target value subgradient method*, Math. Programming **49** (1991) 359–369.

[KiA90] S. Kim and H. Ahn, *Convergence properties of the modified subgradient method of Camerini et al.*, Naval Res. Logist. Quart. **37** (1990) 961–966.

[KiU89] S. Kim and B.-S. Um, *Polyak's subgradient method with simplified projection for nondifferentiable optimization with linear constraints*, Optimization **20** (1989) 451–456.

[Kiw85] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133, Springer-Verlag, Berlin, 1985.

[Kiw89] ——, *A survey of bundle methods for nondifferentiable optimization*, in Mathematical Programming: Recent Developments and Applications, M. Iri and K. Tanabe, eds., KTT/Kluwer, Tokyo, 1989, pp. 263–282.

[Kiw90] ——, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming **46** (1990) 105–122.

[Kiw92] ——, *Block-iterative surrogate projection methods for convex feasibility problems*, Tech. report, Systems Research Institute, Warsaw, 1992.

[KKA87] S. Kim, S. Koh and H. Ahn, *Two-direction subgradient method for non-differentiable optimization problems*, Oper. Res. Lett. **6**, no. 1 (1987) 43–46.

[KuF90] A. N. Kulikov and V. R. Fazilov, *Convex optimization with prescribed accuracy*, Zh. Vychisl. Mat. i Mat. Fiz. **30**, no. 5 (1990) 663–671 (Russian).

[Lem89] C. Lemaréchal, *Nondifferentiable optimization*, in Optimization, G. L. Nemhauser, A. H. G. Rinnooy-Kan and M. J. Todd, eds., Handbooks for Operations Research and Management Science, vol. 1, North-Holland, Amsterdam, 1989, pp. 529–572.

[LNN91] C. Lemaréchal, A. S. Nemirovskii and Yu. E. Nesterov, *New variants of bundle methods*, Research Report 1508, INRIA, Rocquencourt, 1991.

[LoH88] G. Lou and S.-P. Han, *A parallel projection method for solving generalized linear least-squares problems*, Numer. Math. **53** (1988) 255–264.

[MoS54] T. Motzkin and I. J. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math. 6 (1954) 393–404.

[MTA81] J. F. Maurras, K. Trumper and M. Akgül, *Polynomial algorithms for a class of linear programs*, Math. Programming **21** (1981) 121–136.

[NeY79] A. S. Nemirovskii and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Nauka, Moscow, 1979 (Russian). English transl. Wiley, New York, 1983.

[Oko92] S. O. Oko, *Surrogate methods for linear inequalities*, J. Optim. Theory Appl. **72** (1992) 247–268.

[Pol69] B. T. Polyak, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz. **9**, no. 3 (1969) 509–521 (Russian). English transl. in U.S.S.R. Comput. Math. and Math. Phys. 9 (1969) 14–29.

[SaK87] S. Sarin and M. H. Karwan, *Computational evaluation of two subgradient search methods*, Comput. Oper. Res. **14**, no. 3 (1987) 241–247.

[ScZ92] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim. **2** (1992) 121–152.

[SeS86] S. Sen and H. D. Sherali, *A class of convergent primal-dual subgradient algorithms for decomposable convex programs*, Math. Programming **35**, no. 3 (1986) 279–297.

[Shc79] M. B. Shchepakin, *On a modification of a class of algorithms for mathematical programming*, Zh. Vychisl. Mat. i Mat. Fiz. **19**, no. 6 (1979) 1387–1295 (Russian).

[Shc87] ———, *On the method of orthogonal descent*, Kibernetika no. 1 (1987) 58–62 (Russian).

[Shc92] ———, *An algorithm of orthogonal descent for searching for the zero of a convex function, identification of unsolvability of the problem and the questions of the speed of convergence*, Kibernetika no. 5 (1992) 87–96 (Russian).

[ShM88] H. D. Sherali and D. C. Myers, *Dual formulations and subgradient optimization strategies for linear programming relaxations of mixed-integer programs*, Discrete Appl. Math. **20** (1988) 51–68.

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

[ShU89] H. D. Sherali and O. Ulular, *A primal-dual conjugate subgradient algorithm for specially structured linear and convex programming problems*, Appl. Math. Optim. **20** (1989) 193–221.

[SKR87] S. Sarin, M. H. Karwan and R. L. Radin, *A new surrogate-dual multiplier search procedure*, Naval Res. Logist. Quart. **34** (1987) 431–450.

[Spi87] J. E. Spingarn, *A projection method for least-squared solutions to overdetermined systems of linear inequalities*, Linear Algebra Appl. **86** (1987) 211–236.

[Tel82] J. Telgen, *On relaxation methods for systems of linear inequalities*, European J. Oper. Res. **9** (1982) 184–189.

[Tod79] M. J. Todd, *Some remarks on the relaxation method for linear inequalities*, Tech. Report 419, Cornell Univ., Cornell, Ithaca, 1979.

[Tse90] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming **48** (1990) 249–263.

[YaM92] K. Yang and K. G. Murty, *New iterative methods for linear inequalities*, J. Optim. Theory Appl. **72** (1992) 163–185.

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LÈS NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399