An Infeasible-Point Subgradient Method Using Adaptive Approximate Projections*

Dirk A. Lorenz[1], Marc E. Pfetsch[2], and Andreas M. Tillmann[2]

[1] Institute for Analysis and Algebra, TU Braunschweig, Germany
[2] Research Group Optimization, TU Darmstadt, Germany

Abstract. We propose a new subgradient method for the minimization of nonsmooth convex functions over a convex set. To speed up computations we use adaptive approximate projections only requiring to move within a certain distance of the exact projections (which decreases in the course of the algorithm). In particular, the iterates in our method can be infeasible throughout the whole procedure. Nevertheless, we provide conditions which ensure convergence to an optimal feasible point under suitable assumptions. One convergence result deals with step size sequences that are fixed a priori. Two other results handle dynamic Polyak-type step sizes depending on a lower or upper estimate of the optimal objective function value, respectively. Additionally, we briefly sketch two applications: Optimization with convex chance constraints, and finding the minimum ℓ_1-norm solution to an underdetermined linear system, an important problem in Compressed Sensing.

March 8, 2013

1 Introduction

The projected subgradient method [49] is a classical algorithm for the minimization of a nonsmooth convex function f over a convex closed constraint set X, i.e., for the problem

$$\min\ f(x) \quad \text{s.t.} \quad x \in X. \tag{1}$$

One iteration consists of taking a step of size α_k along the negative direction of an arbitrary subgradient h^k of the objective function f at the current point x^k and then computing the next iterate by projection ($\mathcal{P}_X$) onto the feasible set X:

$$x^{k+1} = \mathcal{P}_X(x^k - \alpha_k\, h^k).$$

Over the past decades, numerous extensions and specializations of this scheme have been developed and proven to converge to a minimum (or minimizer). Well-known disadvantages of the subgradient method are its slow local convergence

* This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project "Sparse Exact and Approximate Recovery" under grants LO 1436/3-1 and PF 709/1-1. Moreover, D. Lorenz acknowledges support from the DFG project "Sparsity and Compressed Sensing in Inverse Problems" under grant LO 1436/2-1.

and the necessity to extensively tune algorithmic parameters in order to obtain practical convergence. On the positive side, subgradient methods involve fast iterations and are easy to implement. In fact, they have been widely used in applications and (still) form one of the most popular algorithms for nonsmooth convex minimization.

The main effort in each iteration of the projected subgradient algorithm usually lies in the computation of the projection $\mathcal{P}_X$. Since the projection is the solution of a (smooth) convex program itself, the required time depends on the structure of X and corresponding specialized algorithms. Examples admitting a fast projection include the case where X is the nonnegative orthant or the ℓ_1-norm-ball $\{ x \mid \|x\|_1 \leq \tau \}$, onto which any $x \in \mathbb{R}^n$ can be projected in $\mathcal{O}(n)$ time, see [50]. The projection is more involved if X is, for instance, an affine space or a (convex) polyhedron. In these latter cases, it makes sense to replace the exact projection $\mathcal{P}_X$ by an approximation $\mathcal{P}_X^\varepsilon$. That is, we do not approximate the projection operator uniformly, but, for a given x, we approximate the projected point adaptively up to a desired accuracy. This is formalized by computing points $\mathcal{P}_X^\varepsilon(x)$ with the property that $\|\mathcal{P}_X^\varepsilon(x) - \mathcal{P}_X(x)\| \leq \varepsilon$ for every $\varepsilon \geq 0$. This concept of an absolute accuracy of the projected point is similar in spirit to the adaptive evaluation of operators as, e.g., used in adaptive wavelet methods (cf. the **APPLY**-routine in [13]). Algorithmically, the idea is that during the early phases of the algorithm we do not need a highly accurate projection, and $\mathcal{P}_X^\varepsilon(x)$ can be faster to compute if ε is larger. In the later phases, one then adaptively tightens the requirement on the accuracy.

One particularly attractive situation in which the approach works is the case where X is an affine space, i.e., defined by a linear equation system. Then one can use a truncated iterative method, e.g., a conjugate gradient (CG) approach, to obtain an adaptive approximate projection. We have observed that often only a few steps (2 or 3) of the CG-procedure are needed to obtain a practically convergent method.

In this paper, we focus on the investigation of convergence properties of a general variant of the projected subgradient method which relies on such adaptive approximate projections. We study conditions on the step sizes and on the accuracy requirements ε_k (in each iteration k) in order to achieve convergence of the sequence of iterates to an optimal point, or at least convergence of the function values to the optimum. We investigate two variants of the algorithm. In the first one, the sequence (α_k) of step sizes forms a divergent but square-summable series ($\sum \alpha_k = \infty$, $\sum \alpha_k^2 < \infty$) and is given a priori. The second variant uses dynamic step sizes which depend on the difference of the current function value to a constant *target value* that estimates the optimal value.

A crucial difference of the resulting algorithms to the standard method is the fact that iterates can be infeasible, i.e., are not necessarily contained in X. We thus call the algorithm of this paper *infeasible-point subgradient algorithm* (ISA). As a consequence, the objective function values of the iterates might be smaller than the optimum, which requires a non-standard analysis; see the proofs in Section 3 for details. Moreover, we always assume that X is strictly contained in the interior of the domain $\operatorname{dom} f$ of f. Note that this excludes the case $X = \operatorname{dom} f$, where our algorithm cannot be applied. Furthermore, we assume that every iterate lies in $\operatorname{dom} f$, since otherwise no first-order information is available. This is automatically fulfilled if $\operatorname{dom} f$ is the whole space, or it can be ensured by requiring that the accuracies ε_k are small enough; cf. also Part 4 of Remark 3.

This paper is organized as follows. We first discuss related approaches in the literature. Then we fix some notation and recall a few basics. In the main part of this paper (Sections 2 and 3), we state our infeasible-point subgradient algorithm (ISA) and provide proofs of convergence. In the subsequent sections we briefly discuss some variants of ISA, an example for the adaptive approximate projection operator from the context of convex chance constraints, and an application of ISA to the problem of finding finding the minimum ℓ_1-norm solution of an underdetermined linear equation system, a problem that lately received a lot of attention in the context of compressed sensing (see, e.g., [17, 10, 15]). We finish with some concluding remarks and give pointers to possible extensions as well as topics of future research.

1.1 Related work

The objective function values of the iterates in subgradient algorithms typically do not decrease monotonically. With the right choice of step sizes, the (projected) subgradient method nevertheless guarantees convergence of the objective function values to the minimum, see, e.g., [49, 44, 5, 46]. A typical result of this sort holds for step size sequences (α_k) which are nonsummable ($\sum_{k=0}^{\infty} \alpha_k = \infty$), but square-summable ($\sum_{k=0}^{\infty} \alpha_k^2 < \infty$). Thus, $\alpha_k \to 0$ as $k \to \infty$. Often, the corresponding sequence of points can also be guaranteed to converge to an optimal solution x^*, although this is not necessarily the case; see [3] for a discussion.

Another widely used step size rule uses an estimate φ of the optimal value f^*, a subgradient h^k of the objective function f at the current iterate x^k, and relaxation parameters $\lambda_k > 0$:

$$\alpha_k = \lambda_k \frac{f(x^k) - \varphi}{\|h^k\|_2^2}. \tag{2}$$

The parameters λ_k are constant or required to obey certain conditions needed for convergence proofs. The dynamic rule (2) is a straightforward generalization of the so-called Polyak-type step size rule, which uses $\varphi = f^*$, to the more practical case when f^* is unknown. The convergence results given in [2] extend the work of Polyak [44, 45] to $\varphi \geq f^*$ and $\varphi < f^*$ by imposing certain conditions on the sequence (λ_k). We will generalize these results further, using an adaptive approximate projection operator instead of the (exact) Euclidean projection.

Many extensions of the basic subgradient scheme exist, such as variable target value methods (see, e.g., [14, 28, 36, 40, 48, 19, 5]), using approximate subgradients [6, 1, 34, 16], or incremental projection schemes [23, 40, 31], to name just a few.

Inexact projections have been used previously, probably most prominently for convex feasibility problems in the framework of successive projection methods. Indeed, the optimization problem (1) can, at least theoretically, be cast as the convex feasibility problem to determine $x^* \in X \cap \{f(x) \leq f^*\}$. Using so-called subgradient projections [4] onto the second set leads to a subgradient step

$$x^{k+1} := x^k - \frac{f(x^k) - f^*}{\|h^k\|^2} h^k,$$

which corresponds to using a Polyak-type step size without relaxation parameter, employing the exact optimal value. As illustrated in [4], this approach leads to a very flexible framework for convex feasibility problems as well as (non-smooth) convex optimization problems.

Moreover, [52] considers additive vanishing non-summable error terms (for both the projection and the subgradient step) and establishes the existence of a (decaying) bound on the error terms such that the algorithm will reach a small neighborhood of the optimal set. However, these bounds are not given explicitly. In contrast, our results (Theorems 1 and 3) contain explicit conditions for the error terms that guarantee convergence to the optimum. Another example for the use of inexact projections is the level set subgradient algorithm in [30], although there, all iterates are strictly feasible.

We emphasize that there are at least three conceptually different approaches to approximate projections in the present context. The first concept—prominent, e.g., in the field on convex feasibility problems—uses the idea of approximating the *direction towards the feasible set*, i.e., the iterates approximately move towards the constraint set. In the second, related, concept one *projects exactly onto supersets* of the constraint which are easier to handle, e.g., half-spaces. With both ideas one can use powerful notions like Fejér-monotonicity or the concept of firmly non-expansive mappings, see, e.g., [4] and the more recent [35]; see also the "feasibility operator" framework proposed in [23]. To employ either approach one exploits analytical knowledge about the feasible set, e.g., that it can be written as a level set of a known and easy-to-handle convex function. In the third approach, one aims at *approximating the projected point* without further restricting the direction. This concept applies, for instance, in situations in which a computational error is made in the projection step (e.g., as in [52]) or when it is impossible or undesirable to handle the constraints analytically, but a numerical algorithm is available which calculates the projection point up to a given accuracy. The adaptive approximate projections considered in this paper fall under this third category.

Note that, besides the different philosophies and fields of application, none of the approaches directly dominates the other: On the one hand, one may move directly towards the feasible set while missing the projection point, and on the other hand, one may also move closer to the projected point along a direction which is not towards the feasible set; see Figure 1 for an illustration. However, one can sometimes, for a given rule which approximates the projection direction, find appropriate half-spaces which contain the feasible set and realize

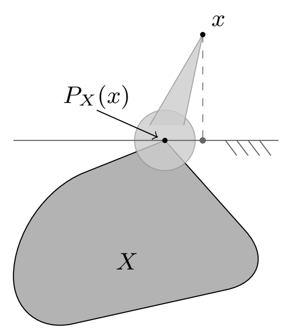


Fig. 1. Schematic illustration of the three concepts of "approximate projections": The approximation of the projection direction (or "moving towards the feasible set") moves from x along a direction within the shaded cone. The exact projection onto a half-space containing X moves along the dashed line. The approximation of the projected point moves from x into a neighborhood of $P_X(x)$, the shaded circle.

this approximate projection exactly. In Section 5 we give a concrete example in which the Fejér-type feasibility operator of [23] is not applicable, but the exact projection point can be approximated reasonably well in the sense of our adaptive approximate projection (see above or (7) below).

In the present paper we only consider the third approach to approximate projections and do not use any assumption like non-expansiveness or Fejér-monotonicity for the iteration mapping in our convergence analyses.

1.2 Notation

In this paper, we consider the convex optimization problem (1) in which we assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function (not necessarily differentiable), $\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$, and $X \subset \operatorname{int}(\operatorname{dom} f) \subseteq \mathbb{R}^n$ is a closed convex set (note that this implies that f is continuous on X). The set

$$\partial f(x) \coloneqq \{h \in \mathbb{R}^n \mid f(y) \geq f(x) + h^\top (y - x) \quad \forall\, y \in \mathbb{R}^n\} \tag{3}$$

is the *subdifferential* of f at a point $x \in \mathbb{R}^n$; its members are the corresponding *subgradients*. Throughout this paper, we will assume (1) to have a nonempty set of optima

$$X^* \coloneqq \operatorname{argmin}\{f(x) \mid x \in X\}. \tag{4}$$

An optimal point will be denoted by x^* and its objective function value $f(x^*)$ by f^*. For a sequence $(x^k) = (x^0, x^1, x^2, \dots)$ of points, the corresponding sequence of objective function values will be abbreviated by $(f_k) = (f(x^k))$.

By $\|\cdot\|_p$ we denote the usual ℓ_p-norm, i.e., for $x \in \mathbb{R}^n$,

$$\|x\|_p \coloneqq \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, & \text{if } 1 \leq p < \infty, \\ \max\limits_{i=1,\dots,n} |x_i|, & \text{if } p = \infty. \end{cases} \tag{5}$$

If no confusion can arise, we shall simply write $\|\cdot\|$ instead of $\|\cdot\|_2$ for the Euclidean (ℓ_2-)norm. The Euclidean distance of a point x to a set Y is

$$d_Y(x) := \inf_{y \in Y} \|x - y\|_2. \tag{6}$$

For Y closed and convex, (6) has a unique minimizer, namely the orthogonal (Euclidean) projection of x onto Y, denoted by $\mathcal{P}_Y(x)$.

All further notation will be introduced where it is needed.

2 The Infeasible-Point Subgradient Algorithm (ISA)

In the projected subgradient algorithm, we replace the exact projection $\mathcal{P}_X$ by an adaptive approximate projection. We require that we can adapt the accuracy of the approximation of the projected point absolutely, i.e., that for any given accuracy parameter $\varepsilon \geq 0$, the adaptive approximate projection $\mathcal{P}_X^\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ satisfies

$$\|\mathcal{P}_X^\varepsilon(x) - \mathcal{P}_X(x)\| \leq \varepsilon \qquad \text{for all } x \in \mathbb{R}^n. \tag{7}$$

In particular, for $\varepsilon = 0$, we have $\mathcal{P}_X^0 = \mathcal{P}_X$. Note that $\mathcal{P}_X^\varepsilon(x)$ does not necessarily produce a point that is *closer* to $\mathcal{P}_X(x)$ (or even to X) than x itself. In fact, this is only guaranteed for $\varepsilon < d_X(x)$.

One example arises in the context of convex chance constraints and is discussed in Section 5.1. For the special case in which X is an affine space, we give a detailed discussion of an adaptive approximate projection satisfying the above requirement in Section 5.2.

By replacing the exact by an adaptive projection in the projected subgradient method, we obtain the *Infeasible-point Subgradient Algorithm* (ISA), which we will discuss in two variants in the following.

The stopping criteria of the algorithms will be ignored for the convergence analyses. In practical implementations, one would stop, e.g., if no significant progress in the objective (or feasibility) has occurred within a certain number of iterations.

2.1 ISA with a predetermined step size sequence

If the step sizes (α_k) and projection accuracies (ε_k) are *predetermined* (i.e., given a priori), we obtain Algorithm 1. Note that $h^k = 0$ might occur, but does not necessarily imply that x^k is optimal, because x^k may be infeasible. In such a case, the adaptive projection will change x^k to a different point as soon as ε_k becomes small enough.

We will now state our main convergence result for this variant of the ISA, using fairly standard step size conditions. The proof is provided in Section 3.

Theorem 1 (Convergence for predetermined step size sequences). *Let the projection accuracy sequence* (ε_k) *be such that*

$$\varepsilon_k \geq 0, \quad \sum_{k=0}^{\infty} \varepsilon_k < \infty, \tag{8}$$

Algorithm 1 PREDETERMINED STEP SIZE ISA

Input: a starting point x^0, sequences (α_k), (ε_k)
Output: an (approximate) solution to (1)
1: initialize $k := 0$
2: **repeat**
3: choose a subgradient $h^k \in \partial f(x^k)$ of f at x^k
4: compute the next iterate $x^{k+1} := \mathcal{P}_X^{\varepsilon_k}\left(x^k - \alpha_k h^k\right)$
5: increment $k := k + 1$
6: **until** a stopping criterion is satisfied

let the positive step size sequence (α_k) *be such that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \tag{9}$$

and let the following relation hold:

$$\alpha_k \geq \sum_{j=k}^{\infty} \varepsilon_j \qquad \forall\, k = 0, 1, 2, \dots \tag{10}$$

Suppose $\|h^k\| \leq H < \infty$ *for all* k*. Then the sequence of the ISA iterates* (x^k) *converges to an optimal point.*

Remark 1. Relations (8), (9), and (10) can be ensured, e.g., by the sequences $\varepsilon_k = 1/k^2$ and $\alpha_k = 1/(k-1)$ for $k > 1$; in particular,

$$\sum_{j=k}^{\infty} \varepsilon_k \leq \int_{k-1}^{\infty} \frac{1}{x^2}\, dx = \frac{1}{k-1} = \alpha_k.$$

2.2 ISA with dynamic step sizes

In order to apply the dynamic step size rule (2), we need several modifications of the basic method, yielding Algorithm 2. This algorithm works with an estimate φ of the optimal objective function value f^* and essentially tries to reach a feasible point x^k with $f(x^k) \leq \varphi$. (Note that if $\varphi = f^*$, we would have obtained an optimal point in this case.)

Remark 2. A few comments on Algorithm 2 are in order:

1. Since $0 < \gamma < 1$, $\gamma^\ell \to 0$ (strictly monotonically) for $\ell \to \infty$. Thus, Steps 3–7 constitute a *projection accuracy refinement phase*, i.e., an inner loop in which the current k is temporarily fixed, and x^k is *recomputed* with a stricter accuracy setting for the adaptive projection. This phase either leads to a point showing $\varphi \geq f^*$ (by termination or convergence in the inner loop over ℓ) or eventually resets x^k to a point with $f_k > \varphi$ and $h^k \neq 0$ so that the regular (outer) iteration is resumed (with k no longer fixed).

Algorithm 2 Dynamic Step Size ISA

Input: estimate φ of f^*, starting point x^0, sequences (λ_k), (ε_k), parameter $\gamma \in (0,1)$
Output: an (approximate) solution to (1)

1: initialize $k := 0$, $\ell = -1$, $x^{-1} := x^0$, $h^{-1} := 0$, $\alpha_{-1} := 0$, $\varepsilon_{-1} := \varepsilon_0$
2: **repeat**
3: choose a subgradient $h^k \in \partial f(x^k)$ of f at x^k
4: **if** $f_k \leq \varphi$ **or** $h^k = 0$ **then**
5: **if** $x^k \in X$ **then**
6: stop (at feasible point x^k showing $\varphi \geq f^*$; optimal if $h^k = 0$)
7: increment $\ell := \ell + 1$, reset $x^k := \mathcal{P}_X^{\varepsilon}(x^{k-1} - \alpha_{k-1} h^{k-1})$ for $\varepsilon = \gamma^\ell \varepsilon_{k-1}$
8: go to Step 3
9: compute step size $\alpha_k := \lambda_k (f_k - \varphi)/\|h^k\|^2$
10: compute the next iterate $x^{k+1} := \mathcal{P}_X^{\varepsilon_k}(x^k - \alpha_k h^k)$
11: reset $\ell := 0$ and increment $k := k + 1$
12: **until** a stopping criterion is satisfied

2. Note that, if x^0 is such that $f_0 \leq \varphi$ or $h^k = 0$, the algorithm begins with such a refinement phase, projecting x^0 more and more accurately until neither case holds any longer (if possible); the initializations with counter -1 are needed for this eventuality. Moreover, we could clearly postpone the (repeated) determination of a subgradient (Step 3) in a refinement phase until $f_k > \varphi$ is achieved, i.e., $h^k = 0$ would be the only reason for another accuracy refinement. This may be important in practice, where finding a subgradient sometimes is expensive itself, and the case $h^k = 0$ presumably occurs very rarely anyway. For the sake of brevity we did not treat this explicitly in Algorithm 2.
3. There are various ways in which the accuracy refinement phase could be realized. Instead of (γ^ℓ) with constant $\gamma \in (0,1)$, any (strictly) monotonically decreasing sequence (γ_ℓ) could be used. Since we will need $\varepsilon_k \to 0$ to achieve feasibility (in the limit) anyway, which implies that for all k there always exists some $L > 0$ such that $\varepsilon_{k+L} < \varepsilon_k$, we could also use $\min\{\varepsilon_{k-1}, \varepsilon_{k-1+\ell}\}$ as the recalibrated accuracy. Moreover, we do not need to fix k, i.e., repeatedly *replace* x^k by finer approximate projections, but could produce a finite series of identical iterates (each reset to the last one before the inner loop started) until the refinement phase is over. Similarly, we could use $\alpha_k = \max\{0, \lambda_k(f_k - \varphi)/\|h^k\|^2\}$ (and 0 if $h^k = 0$); letting $\varepsilon_k \to 0$ then naturally implements the refinement, while in iterations with $\alpha_k = 0$, the produced point may move up to ε_k *away* from the optimal set. Assuming (ε_k) is summable, this does not impede convergence. For all these variants, analogues to the following convergence results hold true as well; however, the proofs require some extensions to account for the technical differences to the variant we chose to present, which admitted the overall shortest proofs. In practice, we would generally expect these variants to behave similarly. Furthermore, note that in principle, the "problematic" cases could also be treated by reverting to exact projections; however, in our present context

this should be avoided since computing the exact projection is considered too expensive.

We obtain the following convergence results, depending on whether φ over- or underestimates f^*. The proofs are deferred to the next section.

Theorem 2 (Convergence for dynamic step sizes with overestimation). *Let the optimal point set X^* be bounded, $\varphi \geq f^*$, $0 < \lambda_k \leq \beta < 2$ for all k, and $\sum_{k=0}^{\infty} \lambda_k = \infty$. Let (ν_k) be a nonnegative sequence with $\sum_{k=0}^{\infty} \nu_k < \infty$, and let*

$$\bar{\varepsilon}_k := -\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right) + \sqrt{\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)^2 + \frac{\lambda_k(2-\lambda_k)(f_k-\varphi)^2}{\|h^k\|^2}}. \tag{11}$$

If the subgradients h^k satisfy $0 < \underline{H} \leq \|h^k\| \leq \overline{H} < \infty$ and (ε_k) satisfies $0 \leq \varepsilon_k \leq \min\{\bar{\varepsilon}_k,\, \nu_k\}$ for all k, then the following holds.

(i) *For any given $\delta > 0$ there exists some index K such that $f(x^K) \leq \varphi + \delta$.*
(ii) *If additionally $f(x^k) > \varphi$ for all k and if $\lambda_k \to 0$, then $f_k \to \varphi$ for $k \to \infty$.*

Remark 3.

1. The sequence (ν_k) is a technicality needed in the proof to ensure $\varepsilon_k \to 0$. Note from (11) that $\bar{\varepsilon}_k > 0$ as long as ISA keeps iterating (in the main loop over k), since $f_k > \varphi$ is then guaranteed by the adaptive accuracy refinements and $0 < \lambda_k < 2$ holds by assumption.
2. More precisely, part (i) of Theorem 2 essentially means that after a finite number of iterations, we reach a point x^k with $f^* - c \leq f(x^k) \leq \varphi + \delta$ for any $c > 0$. If $\varphi < f(x^k) \leq \varphi + \delta$, this point may still be infeasible, but the closer $f(x^k)$ gets to φ, the smaller $\bar{\varepsilon}_k$ becomes, i.e., the algorithm automatically increases the projection accuracy. On the other hand, termination in Step 6 implies that $f(x^k) \geq f^*$ (since x^k is then feasible), and if some inner loop is infinite, then the refined projection points converge to a feasible point. Hence, for every $c > 0$, there is some integer $0 \leq L < \infty$ such that after the L-th accuracy refinement and replacement of x^k, $f(x^k) \geq f^* - c$.
3. Part (ii) shows what happens when all function values $f(x^k)$ stay above the overestimate φ of f^*—which particularly holds true *after* possible refinements, if all the accuracy refinement phases are finite (and no termination occurs)—and we impose $\lambda_k \to 0$ for $k \to \infty$: We eventually obtain $f(x^k)$ arbitrarily close to φ, with vanishing feasibility violation as $k \to \infty$. Then, as well as in case of termination in Step 6 or convergence in a refinement phase ($\ell \to \infty$), it may be desirable to restart the algorithm using a smaller φ; see Section 4.2.
4. The conditions $\|h^k\| \geq \underline{H} > 0$, for all k, in Theorem 2 imply that all subgradients used by the algorithm are nonzero. These conditions are often automatically guaranteed, for example, if X is compact and no unconstrained

optimum of f lies in X. In this case, $\|h\| \geq \underline{H} > 0$ for all $h \in \partial f(x)$ and $x \in X$. Moreover, the same holds for a small enough open neighborhood of X. Also, the norms of the subgradients are bounded from above. Thus, if we start close enough to X and restrict ε_k to be small enough, the conditions of Theorem 2 are fulfilled. Another example in which the conditions are satisfied appears in Section 5.2.

Theorem 3 (Convergence for dynamic step sizes with underestimation). *Let the set of optimal points X^* be bounded, $\varphi < f^*$, $0 < \lambda_k \leq \beta < 2$ for all k, and $\sum_{k=0}^{\infty} \lambda_k = \infty$. Let (ν_k) be a nonnegative sequence with $\sum_{k=0}^{\infty} \nu_k < \infty$, let*

$$L_k := \frac{\lambda_k(2-\beta)(f_k - \varphi)}{\|h^k\|^2}\big(f^* - f_k + \frac{\beta}{2-\beta}(f^* - \varphi)\big), \tag{12}$$

and let

$$\tilde{\varepsilon}_k := -\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right) + \sqrt{\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)^2 - L_k}. \tag{13}$$

If the subgradients h^k satisfy $0 < \underline{H} \leq \|h^k\| \leq \overline{H} < \infty$ and (ε_k) satisfies $0 \leq \varepsilon_k \leq \min\{|\tilde{\varepsilon}_k|, \nu_k\}$ for all k, then the following holds.

(i) *For any given $\delta > 0$, there exists some K such that $f_K \leq f^* + \frac{\beta}{2-\beta}(f^* - \varphi) + \delta$.*
(ii) *If additionally $\lambda_k \to 0$, then the sequence of objective function values (f_k) of the ISA iterates (x^k) converges to the optimal value f^*.*

Remark 4.

1. If $f(x^k) \leq \varphi < f^*$, Steps 3–7 ensure that after a finite number of projection refinements x^k satisfies $\varphi < f(x^k)$. Thus, the algorithm will never terminate with Step 6 and every refinement phase is finite.
2. Moreover, infeasible points x^k with $\varphi < f(x^k) < f^*$ are possible. Hence, the inequality in Theorem 3 (i) may be satisfied too soon to provide conclusive information regarding solution quality. Interestingly, part (ii) shows that by letting the parameters (λ_k) tend to zero, one can nevertheless establish convergence to the optimal value f^* (and $d_X(x^k) \leq d_{X^*}(x^k) \to 0$, i.e., asymptotic feasibility).
3. Theoretically, small values of β yield smaller errors, while in practice this restricts the method to very small steps (since $\lambda_k \leq \beta$), resulting in slow convergence. This illustrates a typical kind of trade-off between solution accuracy and speed.
4. The use of $|\tilde{\varepsilon}_k|$ in Theorem 3 avoids conflicting bounds on ε_k in case $L_k > 0$. Because $0 \leq \varepsilon_k \leq \nu_k$ holds notwithstanding, $0 \leq \varepsilon_k \to 0$ is maintained.
5. The same statements on lower and upper bounds on $\|h^k\|$ as in Remark 3 apply in the context of Theorem 3.

3 Convergence of ISA

From now on, let (x^k) denote the sequence of points with corresponding objective function values (f_k) and subgradients (h^k), $h^k \in \partial f(x^k)$, as generated by ISA in the respective variant under consideration.

Let us consider some basic inequalities which will be essential in establishing our main results. The exact Euclidean projection is nonexpansive, therefore

$$\|\mathcal{P}_X(y) - x\| \leq \|y - x\| \quad \forall x \in X. \tag{14}$$

Hence, for the adaptive approximate projection $\mathcal{P}_X^\varepsilon$ we have, by (7) and (14), for all $x \in X$

$$\begin{aligned}\|\mathcal{P}_X^\varepsilon(y) - x\| &= \|\mathcal{P}_X^\varepsilon(y) - \mathcal{P}_X(y) + \mathcal{P}_X(y) - x\| \\ &\leq \|\mathcal{P}_X^\varepsilon(y) - \mathcal{P}_X(y)\| + \|\mathcal{P}_X(y) - x\| \leq \varepsilon + \|y - x\|.\end{aligned} \tag{15}$$

At some iteration k, let x^{k+1} be produced by ISA using some step size α_k and write $y^k \coloneqq x^k - \alpha_k h^k$. We thus obtain for every $x \in X$:

$$\begin{aligned}&\|x^{k+1} - x\|^2 = \|\mathcal{P}_X^{\varepsilon_k}(y^k) - x\|^2 \\ \leq\;& \big(\|y^k - x\| + \varepsilon_k\big)^2 = \|y^k - x\|^2 + 2\,\|y^k - x\|\,\varepsilon_k + \varepsilon_k^2 \\ =\;& \|x^k - x\|^2 - 2\,\alpha_k (h^k)^\top (x^k - x) + \alpha_k^2\,\|h^k\|^2 + 2\,\|y^k - x\|\,\varepsilon_k + \varepsilon_k^2 \\ \leq\;& \|x^k - x\|^2 - 2\,\alpha_k (f_k - f(x)) + \alpha_k^2\,\|h^k\|^2 + 2\|x^k - x\|\varepsilon_k + 2\,\alpha_k\,\varepsilon_k \|h^k\| + \varepsilon_k^2 \\ =\;& \|x^k - x\|^2 - 2\,\alpha_k (f_k - f(x)) + \big(\alpha_k\,\|h^k\| + \varepsilon_k\big)^2 + 2\,\|x^k - x\|\,\varepsilon_k,\end{aligned} \tag{16}$$

where the second inequality follows from the subgradient definition (3) and the triangle inequality. Note that the above inequalities (14)–(16) hold in particular for every optimal point $x^* \in X^*$.

3.1 ISA with predetermined step size sequence

The proof of the convergence of the ISA iterates x^k is somewhat more involved than for the classical subgradient method as, e.g., in [49]. This is due to the additional error terms by adaptive approximate projection and the fact that $f_k \geq f^*$ is not guaranteed since the iterates may be infeasible.

Proof of Theorem 1. We rewrite the estimate (16) with $x = x^* \in X^*$ as

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\,\alpha_k\,(f_k - f^*) + \underbrace{\big(\alpha_k \|h^k\| + \varepsilon_k\big)^2 + 2\,\|x^k - x^*\|\,\varepsilon_k}_{=:\beta_k} \tag{17}$$

and obtain (by applying (17) for $k = 0, \dots, m$)

$$\|x^{m+1} - x^*\|^2 \;\leq\; \|x^0 - x^*\|^2 - 2\sum_{k=0}^{m} (f_k - f^*)\alpha_k + \sum_{k=0}^{m} \beta_k.$$

Our first goal is to show that $\sum_k \beta_k$ is a convergent series. Using $\|h^k\| \le H$ and denoting $A \coloneqq \sum_{k=0}^{\infty} \alpha_k^2$, we get

$$\sum_{k=0}^{m} \beta_k \ \le \ AH^2 + \sum_{k=0}^{m} \varepsilon_k^2 + 2H \sum_{k=0}^{m} \alpha_k \varepsilon_k + 2 \sum_{k=0}^{m} \|x^k - x^*\| \varepsilon_k.$$

Now denote $D \coloneqq \|x^0 - x^*\|$ and consider the last term (without the factor 2):

$$\begin{aligned}
\sum_{k=0}^{m} \|x^k - x^*\| \varepsilon_k \ &= \ D\,\varepsilon_0 + \sum_{k=1}^{m} \left\| \mathcal{P}_X^{\varepsilon_{k-1}} \left(x^{k-1} - \alpha_{k-1} h^{k-1} \right) - x^* \right\| \varepsilon_k \\
\le \ & D\,\varepsilon_0 + \sum_{k=1}^{m} \left\| \mathcal{P}_X^{\varepsilon_{k-1}} \left(x^{k-1} - \alpha_{k-1} h^{k-1} \right) - \mathcal{P}_X \left(x^{k-1} - \alpha_{k-1} h^{k-1} \right) \right\| \varepsilon_k \\
& \quad + \sum_{k=1}^{m} \left\| \mathcal{P}_X \left(x^{k-1} - \alpha_{k-1} h^{k-1} \right) - x^* \right\| \varepsilon_k \\
\le \ & D\,\varepsilon_0 + \sum_{k=1}^{m} \varepsilon_{k-1} \varepsilon_k + \sum_{k=1}^{m} \left\| x^{k-1} - \alpha_{k-1} h^{k-1} - x^* \right\| \varepsilon_k \\
\le \ & D\,\varepsilon_0 + \sum_{k=0}^{m-1} \varepsilon_k \varepsilon_{k+1} + \sum_{k=0}^{m-1} \|x^k - x^*\|\, \varepsilon_{k+1} + \sum_{k=0}^{m-1} \|h^k\|\, \alpha_k\, \varepsilon_{k+1} \\
\le \ & D\,(\varepsilon_0 + \varepsilon_1) + \sum_{k=0}^{m-1} \varepsilon_k \varepsilon_{k+1} + \sum_{k=1}^{m-1} \|x^k - x^*\|\, \varepsilon_{k+1} + H \sum_{k=0}^{m-1} \alpha_k\, \varepsilon_{k+1}. \qquad (18)
\end{aligned}$$

Repeating this procedure to eliminate all terms $\|x^k - x^*\|$ for $k > 0$, we obtain

$$\begin{aligned}
(18) \ \le \ \dots \ \le \ & D \sum_{k=0}^{m} \varepsilon_k + \sum_{j=1}^{m} \Big(\sum_{k=0}^{m-j} \varepsilon_k \varepsilon_{k+j} + H \sum_{k=0}^{m-j} \alpha_k \varepsilon_{k+j} \Big) \\
= \ & D \sum_{k=0}^{m} \varepsilon_k + \sum_{j=1}^{m} \sum_{k=0}^{m-j} (\varepsilon_k + H\alpha_k)\, \varepsilon_{k+j}. \qquad (19)
\end{aligned}$$

Using the above chain of inequalities, (8) and (10), and the abbreviation $E \coloneqq \sum_{k=0}^{\infty} \varepsilon_k$, we finally get:

$$\begin{aligned}
& \|x^{m+1} - x^*\|^2 + 2 \sum_{k=0}^{m} (f_k - f^*)\, \alpha_k \le D^2 + \sum_{k=0}^{m} \beta_k \\
\le \ & D^2 + AH^2 + \sum_{k=0}^{m} \varepsilon_k^2 + 2\,H \sum_{k=0}^{m} \alpha_k \varepsilon_k + 2\,D \sum_{k=0}^{m} \varepsilon_k + 2 \sum_{j=1}^{m} \sum_{k=0}^{m-j} (\varepsilon_k + H\alpha_k)\, \varepsilon_{k+j} \\
\le \ & D^2 + AH^2 + 2\,D \sum_{k=0}^{m} \varepsilon_k + 2 \sum_{j=0}^{m} \sum_{k=0}^{m-j} \varepsilon_k \varepsilon_{k+j} + 2\,H \sum_{j=0}^{m} \sum_{k=0}^{m-j} \alpha_k \varepsilon_{k+j}
\end{aligned}$$

$$
\begin{aligned}
&= D^2 + AH^2 + 2\,D \sum_{k=0}^{m} \varepsilon_k + 2 \sum_{j=0}^{m} \Big(\varepsilon_j \sum_{k=j}^{m} \varepsilon_k \Big) + 2\,H \sum_{j=0}^{m} \Big(\alpha_j \sum_{k=j}^{m} \varepsilon_k \Big) \\
&\le D^2 + AH^2 + 2\,D \sum_{k=0}^{m} \varepsilon_k + 2 \sum_{j=0}^{m} E\,\varepsilon_j + 2H \sum_{j=0}^{m} \alpha_j\,\alpha_j \\
&\le D^2 + AH^2 + 2\,(D+E) \sum_{k=0}^{m} \varepsilon_k + 2\,H \sum_{k=0}^{m} \alpha_k^2 \\
&\le (D+E)^2 + E^2 + (2+H)\,A\,H \;=:\; R \;<\; \infty. \qquad (20)
\end{aligned}
$$

Since the iterates x^k may be infeasible, possibly $f_k < f^*$, and hence the second term on the left hand side of (20) might be negative. Therefore, we distinguish two cases:

i) If $f_k \ge f^*$ for all but finitely many k, we can assume without loss of generality that $f_k \ge f^*$ for all k (by considering only the "later" iterates). Now, because $f_k \ge f^*$ for all k,

$$
\sum_{k=0}^{m} (f_k - f^*)\,\alpha_k \ge \sum_{k=0}^{m} \Big(\underbrace{\min_{j=0,\dots,m} f_j}_{=:f_m^*} - f^* \Big)\,\alpha_k = (f_m^* - f^*) \sum_{k=0}^{m} \alpha_k.
$$

Together with (20) this yields

$$
0 \le 2\,(f_m^* - f^*) \sum_{k=0}^{m} \alpha_k \le R \quad \Longleftrightarrow \quad 0 \le f_m^* - f^* \le \frac{R}{2 \sum_{k=0}^{m} \alpha_k}.
$$

Thus, because $\sum_{k=0}^{m} \alpha_k$ diverges, we have $f_m^* \to f^*$ for $m \to \infty$ (and, in particular, $\liminf_{k\to\infty} f_k = f^*$).

To show that f^* is in fact the only possible accumulation point (and hence the limit) of (f_k), assume that (f_k) has another accumulation point strictly larger than f^*, say $f^* + \eta$ for some $\eta > 0$. Then, both cases $f_k < f^* + \frac{1}{3}\eta$ and $f_k > f^* + \frac{2}{3}\eta$ must occur infinitely often. We can therefore define two index subsequences (m_ℓ) and (n_ℓ) by setting $n_{(-1)} := -1$ and, for $\ell \ge 0$,

$$
\begin{aligned}
m_\ell &:= \min\{\, k \mid k > n_{\ell-1},\ f_k > f^* + \tfrac{2}{3}\eta \,\}, \\
n_\ell &:= \min\{\, k \mid k > m_\ell,\ f_k < f^* + \tfrac{1}{3}\eta \,\}.
\end{aligned}
$$

Figure 2 illustrates this choice of indices. Now observe that for any ℓ,

$$
\begin{aligned}
\tfrac{1}{3}\eta &< f_{m_\ell} - f_{n_\ell} \le H \cdot \|x^{n_\ell} - x^{m_\ell}\| \le H \left(\|x^{n_\ell - 1} - x^{m_\ell}\| + H\alpha_{n_\ell - 1} + \varepsilon_{n_\ell - 1} \right) \\
&\le \cdots \le H^2 \sum_{j=m_\ell}^{n_\ell - 1} \alpha_j + H \sum_{j=m_\ell}^{n_\ell - 1} \varepsilon_j, \qquad (21)
\end{aligned}
$$

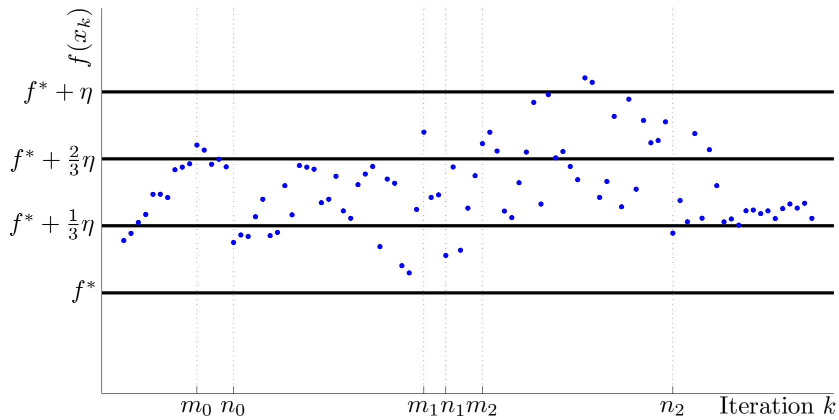


Fig. 2. The sequences (m_ℓ) and (n_ℓ).

where the second inequality is obtained similar to (18). For a given m, let $\ell_m := \max\{\, \ell \mid n_\ell - 1 \le m \,\}$ be the number of blocks of indices between two consecutive indices m_ℓ and $n_\ell - 1$ until m. We obtain:

$$\tfrac{1}{3}\sum_{\ell=0}^{\ell_m} \eta \le H^2 \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \alpha_j + H \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \varepsilon_j \le H^2 \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \alpha_j + HE. \tag{22}$$

For $m \to \infty$, the left hand side tends to infinity, and since $HE < \infty$, this implies that

$$\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \alpha_j \to \infty.$$

Then, since $\alpha_k > 0$ and $f_k \ge f^*$ for all k, (20) yields

$$\begin{aligned} \infty > R &\ge \|x^{m+1} - x^*\|^2 + 2\sum_{k=0}^{m}(f_k - f^*)\alpha_k \ge 2\sum_{k=0}^{m}(f_k - f^*)\alpha_k \\ &\ge 2\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \underbrace{(f_j - f^*)}_{> \frac{1}{3}\eta}\alpha_j > \tfrac{2}{3}\eta \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1} \alpha_j. \end{aligned}$$

But for $m \to \infty$, this yields a contradiction since the sum on the right hand side diverges. Hence, there does not exist an accumulation point strictly larger than f^*, so we can conclude $f_k \to f^*$ as $k \to \infty$, i.e., the whole sequence (f_k) converges to f^*.

We now consider convergence of the sequence (x^k). From (20) we conclude that both terms on the left hand side are bounded independently of m. In particular this means (x^k) is a bounded sequence. Hence, by the Bolzano-Weierstraß Theorem, it has a convergent subsequence (x^{k_i}) with $x^{k_i} \to \overline{x}$

(as $i \to \infty$) for some $\overline{x}$. To show that the full sequence (x^k) converges to $\overline{x}$, take any K and any $k_i < K$ and observe from (17) that

$$\|x^K - \overline{x}\|^2 \le \|x^{k_i} - \overline{x}\|^2 + \sum_{j=k_i}^{K-1} \beta_j.$$

Since $\sum_k \beta_k$ is a convergent series (as seen from the second last line of (20)), the right hand side becomes arbitrarily small for k_i and K large enough. This implies $x^k \to \overline{x}$, and since $\varepsilon_k \to 0$, $f_k \to f^*$, and X^* is closed, $\overline{x} \in X^*$ must hold.

ii) Now consider the case where $f_k < f^*$ occurs infinitely often. We write (f_k^-) for the subsequence of (f_k) with $f_k < f^*$ and (f_k^+) for the subsequence with $f_k \ge f^*$. Clearly $f_k^- \to f^*$. Indeed, the corresponding iterates are asymptotically feasible (since the projection accuracy ε_k tends to zero), and hence f^* is the only possible accumulation point of (f_k^-).

Denoting $M_m^- = \{k \le m \mid f_k < f^*\}$ and $M_m^+ = \{k \le m \mid f_k \ge f^*\}$, we conclude from (20) that

$$\|x^{m+1} - x^*\|^2 + 2 \sum_{k \in M_m^+} (f_k - f^*)\, \alpha_k \le R + 2 \sum_{k \in M_m^-} (f^* - f_k)\, \alpha_k. \qquad (23)$$

Note that each summand is non-negative. To see that the right hand side is bounded independently of m, let $y^{k-1} = x^{k-1} - \alpha_{k-1} h^{k-1}$, and observe that here ($k \in M_m^-$), due to $f_k < f^* \le f(\mathcal{P}_X(y^{k-1}))$, we have

$$\begin{aligned} f^* - f_k &\le f\big(\mathcal{P}_X(y^{k-1})\big) - f\big(\mathcal{P}_X^{\varepsilon_{k-1}}(y^{k-1})\big) \\ &\le (h^{k-1})^\top \big(\mathcal{P}_X(y^{k-1}) - \mathcal{P}_X^{\varepsilon_{k-1}}(y^{k-1})\big) \\ &\le \|h^{k-1}\| \cdot \big\|\mathcal{P}_X(y^{k-1}) - \mathcal{P}_X^{\varepsilon_{k-1}}(y^{k-1})\big\| \le H\varepsilon_{k-1}, \end{aligned}$$

using the subgradient and Cauchy-Schwarz inequalities as well as property (7) of $\mathcal{P}_X^\varepsilon$ and the boundedness of the subgradient norms. From (23), using (9) and (10), we thus obtain

$$\|x^{m+1} - x^*\|^2 + 2 \sum_{k \in M_m^+} (f_k - f^*)\alpha_k \le R + 2\, H \sum_{k \in M_m^-} \alpha_k\, \varepsilon_{k-1}$$

$$\le\; R + 2\, H \sum_{k \in M_m^-} \alpha_k \alpha_{k-1} \le R + 2H \sum_{k=0}^{\infty} \alpha_k \alpha_{k-1} \le R + 4\, A\, H < \infty. \qquad (24)$$

Similar to case i), we conclude that both the sequence (x^k) and the series $\sum_{k \in M_m^+} (f_k - f^*)\, \alpha_k$ are bounded.

It remains to show that $f_k^+ \to f^*$. Assume to the contrary that (f_k^+) has an accumulation point $f^* + \eta$ for $\eta > 0$. Similar to before, we construct index subsequences (m_ℓ) and (p_ℓ) as follows: Set $p_{(-1)} := -1$ and define, for $\ell \ge 0$,

$$\begin{aligned} m_\ell &:= \min\{\, k \in M_\infty^+ \mid k > p_{\ell-1},\; f_k > f^* + \tfrac{2}{3}\eta \,\}, \\ p_\ell &:= \min\{\, k \in M_\infty^- \mid k > m_\ell \,\}. \end{aligned}$$

Then $m_\ell, \dots, p_\ell - 1 \in M_\infty^+$ for all ℓ, and we have

$$\tfrac{2}{3}\eta < f_{m_\ell} - f_{p_\ell} \leq H^2 \sum_{j=m_\ell}^{p_\ell - 1} \alpha_j + H \sum_{j=m_\ell}^{p_\ell - 1} \varepsilon_j.$$

Therefore, with $\ell_m \coloneqq \max\{\,\ell \,|\, p_\ell - 1 \leq m\,\}$ for a given m,

$$\tfrac{2}{3}\sum_{\ell=0}^{\ell_m} \eta \leq H^2 \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{p_\ell-1} \alpha_j + H \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{p_\ell-1} \varepsilon_j \leq H^2 \sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{p_\ell-1} \alpha_j + H\,E.$$

Now the left hand side becomes arbitrarily large as $m \to \infty$, so that also $\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{p_\ell-1} \alpha_j \to \infty$, since $HE < \infty$. Note that because $\alpha_k > 0$ and

$$\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{p_\ell-1} \alpha_j \leq \sum_{k \in M_m^+} \alpha_k,$$

this latter series must diverge as well. As a consequence, f^* is itself an (other) accumulation point of (f_k^+): From (24) we have

$$\begin{aligned} \infty > R + 4AH &\geq 2 \sum_{k\in M_m^+} (f_k - f^*)\alpha_k \\ &\geq \sum_{k \in M_m^+} \big(\underbrace{\min\{\, f_j \,|\, j \in M_m^+,\, j \leq m\,\}}_{=:\hat f_m^*} - f^*\big)\,\alpha_k = (\hat f_m^* - f^*) \sum_{k\in M_m^+} \alpha_k, \end{aligned}$$

and thus

$$0 \leq \hat f_m^* - f^* \leq \frac{R + 4\,A\,H}{\sum_{k\in M_m^+} \alpha_k} \to 0 \qquad \text{as } m \to \infty,$$

since $\sum_{k \in M_m^+} \alpha_k$ diverges. But then, knowing $(\hat f_k^*)$ converges to f^*, we can use (m_ℓ) and another index subsequence (n_ℓ), given by

$$n_\ell \coloneqq \min\{\, k \in M_\infty^+ \,|\, k > m_\ell,\, f_k < f^* + \tfrac{1}{3}\eta\,\},$$

to proceed analogously to case i) to arrive at a contradiction and conclude that no $\eta > 0$ exists such that $f^* + \eta$ is an accumulation point of (f_k^+).

On the other hand, since (x^k) is bounded and f is continuous on a neighborhood of X (recall that for all k, x^k is contained in an ε_k-neighborhood of X), (f_k^+) is bounded. Thus, it must have at least one accumulation point. Since $f_k \geq f^*$ for all $k \in M_\infty^+$, the only possibility left is f^* itself. Hence, f^* is the unique accumulation point (i.e., the limit) of the sequence (f_k^+). As this is also true for (f_m^-), the whole sequence (f_k) converges to f^*.

Finally, convergence of the bounded sequence (x^k) to some $\overline{x} \in X^*$ can now be obtained just like in case i), completing the proof. □

3.2 ISA with dynamic Polyak-type step sizes

Let us now turn to dynamic step sizes. In the rest of this section, α_k will always denote step sizes of the form (2).

Since in subgradient methods the objective function values need not decrease monotonically, the key quantity in convergence proofs usually is the distance to the optimal set X^*. For ISA with dynamic step sizes (Algorithm 2), we have the following result concerning these distances:

Lemma 1. *Let $x^* \in X^*$. For the sequence of ISA iterates (x^k), computed with step sizes $\alpha_k = \lambda_k(f_k - \varphi)/\|h^k\|^2$, it holds that*

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \;\leq\; &\|x^k - x^*\|^2 + \varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + \|x^k - x^*\|\right)\varepsilon_k \\ &+ \frac{\lambda_k(f_k - \varphi)}{\|h^k\|^2}\big(\lambda_k(f_k - \varphi) - 2(f_k - f^*)\big). \end{aligned} \tag{25}$$

In particular, also

$$d_{X^*}(x^{k+1})^2 \leq d_{X^*}(x^k)^2 - 2\,\alpha_k(f_k - f^*) + (\alpha_k\|h^k\| + \varepsilon_k)^2 + 2\,d_{X^*}(x^k)\,\varepsilon_k. \tag{26}$$

Proof. Plug (2) into (16) for $x = x^*$ and rearrange terms to obtain (25). If the optimization problem (1) has a unique optimum x^*, then obviously $\|x^k - x^*\| = d_{X^*}(x^k)$ for all k, so (26) is identical to (25). Otherwise, note that since X^* is the intersection of the closed set X with the level set $\{x \mid f(x) = f^*\}$ of the convex function f, X^* is closed (cf., for example, [26, Prop. 1.2.2, 1.2.6]) and the projection onto X^* is well-defined. Then, considering $x^* = \mathcal{P}_{X^*}(x^k)$, (16) becomes

$$\left\|x^{k+1} - \mathcal{P}_{X^*}(x^k)\right\|^2 \leq d_{X^*}(x^k)^2 - 2\alpha_k(f_k - f^*) + (\alpha_k\|h^k\| + \varepsilon_k)^2 + 2\,d_{X^*}(x^k)\,\varepsilon_k.$$

Furthermore, because obviously $f(\mathcal{P}_{X^*}(x)) = f(\mathcal{P}_{X^*}(y)) = f^*$ for all $x, y \in \mathbb{R}^n$, and by definition of the Euclidean projection,

$$d_{X^*}(x^{k+1})^2 \;=\; \left\|x^{k+1} - \mathcal{P}_{X^*}(x^{k+1})\right\|^2 \;\leq\; \left\|x^{k+1} - \mathcal{P}_{X^*}(x^k)\right\|^2.$$

Combining the last two inequalities yields (26).

Moreover, note that these results continue to hold true if x^{k+1} is replaced in a projection refinement phase (starting in the next iteration $k+1$), since then only accuracy parameters smaller than ε_k are used. □

Typical convergence results are often derived by showing that the sequence $(\|x^k - x^*\|)$ is monotonically decreasing (for arbitrary $x^* \in X^*$) under certain assumptions on the step sizes, subgradients, etc. This is also done in [2], where (25) with $\varepsilon_k = 0$ for all k is the central inequality, cf. [2, Prop. 2]. In our case, i.e., working with adaptive approximate projections as specified by (7), we can follow this principle to derive conditions on the projection accuracies (ε_k) which still allow for a (monotonic) decrease of the distances from the optimal set: If the last summand in (25) is negative, the resulting gap between the distances from X^* of subsequent iterates can be exploited to relax the projection accuracy, i.e., to choose $\varepsilon_k > 0$ without destroying monotonicity.

Naturally, to achieve feasibility (at least in the limit), we will need to have (ε_k) diminishing ($\varepsilon_k \to 0$ as $k \to \infty$). It will become clear that this, combined with summability ($\sum_{k=0}^{\infty} \varepsilon_k < \infty$) and with monotonicity conditions as described above, is already enough to extend the analysis to cover iterations with $f_k < f^*$, which may occur since we project inaccurately.

For different choices of the estimate φ of f^*, we will now derive the proofs of Theorems 2 and 3 via a series of intermediate results. Corresponding results for exact projections ($\varepsilon_k = 0$) can be found in [2]. In fact, our analysis for adaptive approximate projections improves on some of these earlier results (e.g., [2, Prop. 10] states convergence of some *sub*sequence of the function values to the optimum for the case $\varphi < f^*$, whereas Theorem 3 in this paper gives convergence of the whole sequence (f_k), for approximate and also for exact projections).

For the remainder of this section we can assume that ISA does not terminate in Step 6 and that all inner projection accuracy refinement loops are finite. Otherwise, there is some refinement phase starting at iteration $\underline{k}$ such that, as $\ell \to \infty$, $x^{\underline{k}}$ is repeatedly reset to

$$y_{\underline{k}}^{\ell} \coloneqq \mathcal{P}_X^{\gamma^{\ell} \varepsilon_{\underline{k}-1}}(x^{\underline{k}-1} - \alpha_{\underline{k}-1} h^{\underline{k}-1}) \to \mathcal{P}_X^{0}(x^{\underline{k}-1} - \alpha_{\underline{k}-1} h^{\underline{k}-1}) \in X,$$

with $f(y_{\underline{k}}^{\ell}) \to \underline{\varphi} \leq \varphi$; cf. Remarks 3 and 4.

Using overestimates of the optimal value. In this part we will focus on the case $\varphi \geq f^*$. As might be expected, this relation allows for eliminating the unknown f^* from (26).

Lemma 2. *Let $\varphi \geq f^*$ and $\lambda_k \geq 0$. If $f_k \geq \varphi$ for some $k \in \mathbb{N}$, then*

$$d_{X^*}(x^{k+1})^2 \;\leq\; d_{X^*}(x^k)^2 + \varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)\varepsilon_k + \frac{\lambda_k(\lambda_k - 2)(f_k - \varphi)^2}{\|h^k\|^2}. \tag{27}$$

Proof. This follows immediately from Lemma 1, using $f_k \geq \varphi \geq f^*$ and $\lambda_k \geq 0$. □

Note that ISA guarantees $f_k > \varphi$ by sufficiently accurate projection (otherwise the method stops or the inner refinement loop over ℓ, with fixed k, is infinite, indicating φ was too large, see Steps 3-7 of Algorithm 2), and that the last summand in (27) is always negative for $0 < \lambda_k < 2$. Hence, adaptive approximate projections ($\varepsilon_k > 0$) can always be employed without destroying the monotonic decrease of $(d_{X^*}(x^k))$, as long as the ε_k are chosen small enough.

The following result provides a theoretical bound on how large the projection accuracies ε_k may become.

Lemma 3. *Let $0 < \lambda_k < 2$ for all k. For $\varphi \geq f^*$, the sequence $(d_{X^*}(x^k))$ is monotonically decreasing and converges to some $\zeta \geq 0$, if $0 \leq \varepsilon_k \leq \overline{\varepsilon}_k$ for all k, where $\overline{\varepsilon}_k$ is defined in* (11) *of Theorem 2.*

Proof. Considering (27), it suffices to show that for $\varepsilon_k \leq \overline{\varepsilon}_k$, we have

$$\varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)\varepsilon_k + \frac{\lambda_k(\lambda_k - 2)(f_k - \varphi)^2}{\|h^k\|^2} \leq 0. \tag{28}$$

The bound $\overline{\varepsilon}_k$ from (11) is precisely the (unique) positive root of the quadratic function in ε_k given by the left hand side of (28). Thus, we have a monotonically decreasing (i.e., nonincreasing) sequence $(d_{X^*}(x^k))$, and since its members are bounded below by zero, it converges to some nonnegative value, say ζ. □

As a consequence, if X^* is bounded, we obtain boundedness of the iterate sequence (x^k):

Corollary 1. *Let X^* be bounded. If the sequence $(d_{X^*}(x^k))$ is monotonically decreasing, then the sequence (x^k) is bounded.*

Proof. By monotonicity of $(d_{X^*}(x^k))$, making use of the triangle inequality,

$$\begin{aligned} \|x^k\| &= \left\|x^k - \mathcal{P}_{X^*}(x^k) + \mathcal{P}_{X^*}(x^k)\right\| \\ &\leq d_{X^*}(x^k) + \left\|\mathcal{P}_{X^*}(x^k)\right\| \leq d_{X^*}(x^0) + \sup_{x \in X^*} \|x\| < \infty, \end{aligned}$$

since X^* is bounded by assumption. □

We now have all the tools at hand for proving Theorem 2.

Proof of Theorem 2. First, we prove part (i). Let the main assumptions of Theorem 2 hold and suppose—contrary to the desired result (i)—that $f_k > \varphi + \delta$ for all k (possibly after finitely many refinements of the projection accuracy used to compute x^k). By Lemma 2,

$$\begin{aligned} \frac{\lambda_k(2-\lambda_k)(f_k-\varphi)^2}{\|h^k\|^2} &\leq d_{X^*}(x^k)^2 - d_{X^*}(x^{k+1})^2 \\ &\quad + \varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k - \varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)\varepsilon_k. \end{aligned}$$

Since $0 < \underline{H} \leq \|h^k\| \leq \overline{H} < \infty$, $0 < \lambda_k \leq \beta < 2$, and $f_k - \varphi > \delta$ for all k by assumption, we have

$$\frac{\lambda_k(2-\lambda_k)(f_k-\varphi)^2}{\|h^k\|^2} \geq \frac{\lambda_k(2-\beta)\delta^2}{\overline{H}^2}.$$

By Lemma 3, $d_{X^*}(x^k) \leq d_{X^*}(x^0)$. Also, by Corollary 1 there exists $F < \infty$ such that $f_k \leq F$ for all k. Hence, $\lambda_k(f_k - \varphi) \leq \beta(F - \varphi)$, and since $1/\|h^k\| \leq 1/\underline{H}$, we obtain

$$\frac{(2-\beta)\delta^2}{\overline{H}^2}\lambda_k \leq d_{X^*}(x^k)^2 - d_{X^*}(x^{k+1})^2 + \varepsilon_k^2 + 2\left(\frac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0)\right)\varepsilon_k. \tag{29}$$

Summation of the inequalities (29) for $k = 0, 1, \ldots, m$ yields

$$\frac{(2-\beta)\delta^2}{\overline{H}^2} \sum_{k=0}^{m} \lambda_k \leq \; d_{X^*}(x^0)^2 - d_{X^*}(x^{m+1})^2 + \sum_{k=0}^{m} \varepsilon_k^2 + 2\left(\frac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0) \right) \sum_{k=0}^{m} \varepsilon_k.$$

Now, by assumption, the left hand side tends to infinity as $m \to \infty$, while the right hand side remains finite (note that nonnegativity and summability of (ν_k) imply the summability of (ν_k^2), properties that carry over to (ε_k)). Thus, we have reached a contradiction and therefore proven part (i) of Theorem 2, i.e., that $f_K \leq \varphi + \delta$ holds in some iteration K.

We now turn to part (ii): Let the main assumptions of Theorem 2 hold, let $\lambda_k \to 0$ and suppose $f_k > \varphi$ for all k (again, possibly after refinements). Then, since we know from part (i) that the function values fall below every $\varphi + \delta$, we can construct a monotonically decreasing subsequence (f_{K_j}) such that $f_{K_j} \to \varphi$. (To see this, note that if $f_k < \varphi + \delta$ is reached with $f_k < \varphi$, the ensuing refinement phase not necessarily ends with x^k replaced by a point with $\varphi < f_k < \varphi + \delta$, but that then, however, there always exists a $K > k$ such that $\varphi < f_K < \varphi + \delta$, since $\lambda_k \to 0$, $\varepsilon_k \to 0$, and by continuity of f.)

To show that φ is the unique accumulation point of (f_k), assume to the contrary that there is another subsequence of (f_k) which converges to $\varphi + \eta$, with some $\eta > 0$. We can now employ the same technique as in the proof of Theorem 1 to reach a contradiction:

The two cases $f_k < \varphi + \frac{1}{3}\eta$ and $f_k > \varphi + \frac{2}{3}\eta$ must both occur infinitely often, since φ and $\varphi + \eta$ are accumulation points. Set $n_{(-1)} := -1$ and define, for $\ell \geq 0$,

$$m_\ell := \min\{\, k \mid k > n_{\ell-1},\, f_k > \varphi + \tfrac{2}{3}\eta \,\},$$
$$n_\ell := \min\{\, k \mid k > m_\ell,\, f_k < \varphi + \tfrac{1}{3}\eta \,\}.$$

Then, with $\infty > F \geq f_k$ for all k (existing since (x^k) is bounded and therefore so is (f_k)) and the subgradient norm bounds, we obtain

$$\tfrac{1}{3}\eta < f_{m_\ell} - f_{n_\ell} \leq \overline{H}\|x^{m_\ell} - x^{n_\ell}\| \leq \frac{\overline{H}(F-\varphi)}{\underline{H}} \sum_{j=m_\ell}^{n_\ell - 1} \lambda_j + \overline{H} \sum_{j=m_\ell}^{n_\ell - 1} \varepsilon_j$$

and from this, denoting $\ell_m := \max\{\, \ell \mid n_\ell - 1 \leq m \,\}$ for a given m,

$$\tfrac{1}{3} \sum_{\ell=0}^{\ell_m} \eta \leq \frac{\overline{H}(F-\varphi)}{\underline{H}} \sum_{\ell=0}^{\ell_m} \sum_{j=m_\ell}^{n_\ell - 1} \lambda_j + \overline{H} \sum_{\ell=0}^{\ell_m} \sum_{j=m_\ell}^{n_\ell - 1} \varepsilon_j.$$

Since for $m \to \infty$, the left hand side tends to infinity, the same must hold for the right hand side. But since $\sum_{\ell=0}^{\ell_m} \sum_{j=m_\ell}^{n_\ell - 1} \varepsilon_j \leq \sum_{k=0}^{m} \varepsilon_k \leq \sum_{k=0}^{m} \nu_k < \infty$, this implies

$$\sum_{\ell=0}^{\ell_m} \sum_{j=m_\ell}^{n_\ell - 1} \lambda_j \to \infty \qquad \text{for } m \to \infty. \tag{30}$$

Also, using the same estimates as in part (i) above, (27) yields

$$\underbrace{\tfrac{2-\beta}{\overline{H}}}_{=:C_1<\infty}(f_k-\varphi)^2\lambda_k \le d_{X^*}(x^k)^2 - d_{X^*}(x^{k+1})^2 + \varepsilon_k^2 + \underbrace{2\left(\tfrac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0)\right)}_{=:C_2<\infty}\varepsilon_k$$

and thus by summation for $k = 0,\dots,m$ for a given m,

$$C_1\sum_{k=0}^{m}(f_k-\varphi)^2\lambda_k \le d_{X^*}(x^0)^2 - d_{X^*}(x^{m+1})^2 + \sum_{k=0}^{m}\varepsilon_k^2 + C_2\sum_{k=0}^{m}\varepsilon_k. \tag{31}$$

Observe that all summands of the left hand side term are positive, and thus

$$C_1\sum_{k=0}^{m}(f_k-\varphi)^2\lambda_k \ge C_1\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1}\underbrace{(f_j-\varphi)}_{>\frac{1}{3}\eta}{}^2\lambda_j > \frac{C_1\eta^2}{9}\sum_{\ell=0}^{\ell_m}\sum_{j=m_\ell}^{n_\ell-1}\lambda_j.$$

Therefore, as $m \to \infty$, the left hand side of (31) tends to infinity (by (30) and the above inequality) while the right hand side expression remains finite (recall $0 \le \varepsilon_k \le \nu_k$ with (ν_k) summable and thus also square-summable). Thus, we have reached a contradiction, and it follows that φ is the only accumulation point (i.e., the limit) of the whole sequence (f_k).

This proves part (ii) and thus completes the proof of Theorem 2. □

Remark 5. With more technical effort one can argue along the lines of the proof of Theorem 1 to obtain the following result on the convergence of the iterates x^k in the case of Theorem 3: If we additionally assume that $\sum\lambda_k^2 < \infty$ and that $\lambda_k \ge \sum_{j=k}^{\infty}\varepsilon_k$ for all k, then $x^k \to \overline{x}$ for some $\overline{x} \in X$ with $f(\overline{x}) = \varphi$ and $d_{X^*}(\overline{x}) = \zeta \ge 0$ (ζ being the same as in Lemma 3).

Using lower bounds on the optimal value. In the following, we focus on the case $\varphi < f^*$, i.e., using a constant lower bound in the step size definition (2). Such a lower bound is often more readily available than (useful) upper bounds; for instance, it can be computed via the dual problem, or sometimes derived directly from properties of the objective function such as, e.g., nonnegativity of the function values.

Following arguments similar to those in the previous part, we can prove convergence of ISA (under certain assumptions), provided that the projection accuracies (ε_k) obey conditions analogous to those for the case $\varphi \ge f^*$. Moreover, recall that for $\varphi < f^*$, every refinement phase is finite, so that $f_k > \varphi$ is guaranteed for all k; in particular, Step 6 is never executed since $X \cap \{x \mid f(x) < \varphi\} = \emptyset$.

Let us start with analogues of Lemmas 2 and 3.

Lemma 4. *Let $\varphi < f^*$ and $0 < \lambda_k \le \beta < 2$. If $f_k \ge \varphi$ for some $k \in \mathbb{N}$, then*

$$d_{X^*}(x^{k+1})^2 \;\le\; d_{X^*}(x^k)^2 + \varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k-\varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)\varepsilon_k + L_k, \tag{32}$$

where L_k is defined in (12) *of Theorem 3.*

Proof. For $\varphi < f^*$, $0 < \lambda_k \leq \beta < 2$, and $f_k \geq \varphi$, it holds that

$$\lambda_k(f_k-\varphi)-2(f_k-f^*) \;\leq \beta(f_k-\varphi)-2(f_k-f^*) = \beta(f^*-\varphi)+(2-\beta)(f^*-f_k).$$

The claim now follows immediately from Lemma 1. □

Lemma 5. *Let $\varphi < f^*$, let $0 < \lambda_k \leq \beta < 2$ and $f_k \geq f^* + \frac{\beta}{2-\beta}(f^*-\varphi)$ for all k, and let L_k be given by* (12)*. Then $(d_{X^*}(x^k))$ is monotonically decreasing and converges to some $\xi \geq 0$, if $0 \leq \varepsilon_k \leq \tilde{\varepsilon}_k$ for all k, where $\tilde{\varepsilon}_k$ is defined in* (13)*.*

Proof. The condition $f_k \geq f^* + \frac{\beta}{2-\beta}(f^*-\varphi)$ implies $L_k \leq 0$ and hence ensures that adaptive approximate projection can be used while still allowing for a decrease in the distances of the subsequent iterates from X^*. The rest of the proof is completely analogous to that of Lemma 3, considering (32) and (12) to derive the upper bound $\tilde{\varepsilon}_k$ given by (13) on the projection accuracy. □

We can now state the proof of our convergence results for the case $\varphi < f^*$.

Proof of Theorem 3. Let the assumptions of Theorem 3 hold. We start with proving part (i): Let some $\delta > 0$ be given and suppose—contrary to the desired result (i)—that $f_k > f^* + \frac{\beta}{2-\beta}(f^*-\varphi) + \delta$ for all k (possibly after refinements). By Lemma 4,

$$d_{X^*}(x^{k+1})^2 \;\leq\; d_{X^*}(x^k)^2 + \varepsilon_k^2 + 2\left(\frac{\lambda_k(f_k-\varphi)}{\|h^k\|} + d_{X^*}(x^k)\right)\varepsilon_k + L_k.$$

Since $0 < \underline{H} \leq \|h^k\| \leq \overline{H}$, $0 < \lambda_k \leq \beta < 2$, and $\varphi < f_k$, and due to our assumption on f_k, i.e.,

$$f^* - f_k + \tfrac{\beta}{2-\beta}(f^*-\varphi) \;<\; -\,\delta \qquad \text{for all } k,$$

it follows that

$$L_k \;<\; -\,\frac{\lambda_k(2-\beta)(f_k-\varphi)\delta}{\overline{H}^2} \;<\; 0.$$

By Lemma 5, $d_{X^*}(x^k) \leq d_{X^*}(x^0)$, and Corollary 1 again ensures existence of some $F < \infty$ such that $f_k \leq F$ for all k. Because also $\lambda_k(f_k-\varphi) \leq \beta(F-\varphi)$ and $1/\|h^k\| \leq 1/\underline{H}$, we hence obtain

$$\begin{aligned}\frac{\lambda_k(2-\beta)(f_k-\varphi)\delta}{\overline{H}^2} \;<\; -\,L_k \;&\leq\; d_{X^*}(x^k)^2 - d_{X^*}(x^{k+1})^2\\ &\quad + \varepsilon_k^2 + 2\left(\frac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0)\right)\varepsilon_k. \qquad (33)\end{aligned}$$

Summation of these inequalities for $k = 0, 1, \dots, m$ yields

$$\begin{aligned}\frac{(2-\beta)\delta}{\overline{H}^2}\sum_{k=0}^{m}(f_k-\varphi)\lambda_k \;&<\; d_{X^*}(x^0)^2 - d_{X^*}(x^{m+1})^2\\ &\quad + \sum_{k=0}^{m}\varepsilon_k^2 + 2\left(\frac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0)\right)\sum_{k=0}^{m}\varepsilon_k. \qquad (34)\end{aligned}$$

Moreover, our assumption on f_k yields

$$f_k - \varphi \ > \ f^* + \tfrac{\beta}{2-\beta} f^* - \tfrac{\beta}{2-\beta}\varphi + \delta - \varphi \ = \ \tfrac{2}{2-\beta}(f^* - \varphi) + \delta.$$

It follows from (34) that

$$\frac{\big(2(f^*-\varphi)+(2-\beta)\delta\big)\delta}{\overline{H}^2}\sum_{k=0}^{m}\lambda_k < d_{X^*}(x^0)^2 - d_{X^*}(x^{m+1})^2 + \sum_{k=0}^{m}\varepsilon_k^2 + 2\left(\frac{\beta(F-\varphi)}{\underline{H}} + d_{X^*}(x^0)\right)\sum_{k=0}^{m}\varepsilon_k.$$

Now, by assumption, the left hand side tends to infinity as $m \to \infty$, whereas by Lemma 5 and the choice of $0 \le \varepsilon_k \le \min\{|\tilde{\varepsilon}_k|, \nu_k\}$ with a nonnegative summable (and hence also square-summable) sequence (ν_k), the right hand side remains finite. Thus, we have reached a contradiction, and part (i) is proven, i.e., there does exist some K such that $f_K \le f^* + \frac{\beta}{2-\beta}(f^*-\varphi)+\delta$ (after possible refinements of the projection accuracy used to recompute x^K).

Let us now turn to part (ii): Again, let the main assumptions of Theorem 3 hold and let $\lambda_k \to 0$. Recall that for $\varphi < f^*$, we have $f_k > \varphi$ for all k by construction of ISA (refinement loops). We distinguish three cases:

If $f_k < f^*$ holds for all $k \ge k_0$ for some k_0, then $f_k \to f^*$ is obtained immediately, just like in the proof of Theorem 1, by asymptotic feasibility.

On the other hand, if $f_k \ge f^*$ for all k larger than some k_0, then repeated application of part (i) yields a subsequence of (f_k) which converges to f^*: For any $\delta > 0$ we can find an index K such that $f^* \le f_K \le f^* + \frac{\beta}{2-\beta}(f^* - \varphi) + \delta$. Obviously, we get arbitrarily close to f^* if we choose β and δ small enough. However, we have the restriction $\lambda_k \le \beta$. But since $\lambda_k \to 0$, we may "restart" our argumentation if λ_k is small enough and replace β with a smaller one. With the convergent subsequence thus constructed, we can then use the same technique as in the proof of Theorem 2 (ii) to show that (f_k) has no other accumulation point but f^*, whence $f_k \to f^*$ follows.

Finally, when both cases $f_k < f^*$ and $f_k \ge f^*$ occur infinitely often, we can proceed similar to the proof of Theorem 1: The subsequence of function values below f^* converges to f^*, since $\varepsilon_k \to 0$. For the function values greater or equal to f^*, we assume that there is an accumulation point $f^* + \eta$ larger than f^*, deduce that an appropriate sub-sum of the λ_k's diverge and then sum up equation (33) for the respective indices (belonging to $\{k \mid f_k \ge f^*\}$) to arrive at a contradiction. Note that the iterate sequence (x^k) is bounded, due to Corollary 1 (for iterations k with $f_k \ge f^*$) and since the iterates with $\varphi < f_k < f^*$ stay within a bounded neighborhood of the bounded set X^*, since ε_k tends to zero and is summable. Therefore, as f is continuous on a neighborhood of X (which contains all x^k from some k on), (f_k) is bounded as well and therefore must have at least one accumulation point. The only possibility left now is f^*, so we conclude $f_k \to f^*$. □

Remark 6. With $f_k \to f^*$ and $\varepsilon_k \to 0$, we obviously have $d_{X^*}(x^k) \to 0$ in the setting of Theorem 3. Furthermore, Remark 5 applies similarly: With more conditions on λ_k and more technical effort one can obtain convergence of the sequence (x^k) to some $\overline{x} \in X^*$.

4 Discussion

In this section, we will discuss extensions of ISA. We will also illustrate how to obtain bounds on the projection accuracies that are independent of the (generally unknown) distance from the optimal set, and thus computable.

4.1 Extension to ϵ-subgradients

It is noteworthy that the above convergence analyses also work when replacing the subgradients by ϵ-subgradients [6], i.e., replacing $\partial f(x^k)$ by

$$\partial_{\rho_k} f(x^k) := \{\, h \in \mathbb{R}^n \mid f(x) - f(x^k) \geq h^\top (x - x^k) - \rho_k \quad \forall\, x \in \mathbb{R}^n \,\}. \tag{35}$$

(To avoid confusion with the projection accuracy parameters ε_k, we use ρ_k.) For instance, we immediately obtain the following result:

Corollary 2. *Let ISA (Algorithm 1) choose $h^k \in \partial_{\rho_k} f(x^k)$ with $\rho_k \geq 0$ for all k. Under the assumptions of Theorem 1, if (ρ_k) is chosen summable ($\sum_{k=0}^{\infty} \rho_k < \infty$) and such that*

(i) $\rho_k \leq \mu\, \alpha_k$ *for some* $\mu > 0$, *or*
(ii) $\rho_k \leq \mu\, \varepsilon_k$ *for some* $\mu > 0$,

then the sequence of ISA iterates (x^k) converges to an optimal point.

Proof. The proof is analogous to that of Theorem 1; we will therefore only sketch the necessary modifications: Choosing $h^k \in \partial_{\rho_k} f(x^k)$ (instead of $h^k \in \partial f(x^k)$) adds the term $+2\alpha_k \rho_k$ to the right hand side of (16). If $\rho_k \leq \mu\, \alpha_k$ for some constant $\mu > 0$, the square-summability of (α_k) suffices: By upper bounding $2\alpha_k \rho_k$, the constant term $+2\mu A$ is added to the definition of R in (20). Similarly, $\rho_k \leq \mu\, \varepsilon_k$ does not impair convergence under the assumptions of Theorem 1, because then the additional summand in (20) is

$$2 \sum_{k=0}^{m} \alpha_k \rho_k \;\leq\; 2\mu \sum_{k=0}^{m} \alpha_k \varepsilon_k \;\leq\; 2\mu \sum_{k=0}^{m} \Big(\alpha_k \sum_{\ell=k}^{\infty} \varepsilon_k \Big) \;\leq\; 2\mu \sum_{k=0}^{m} \alpha_k^2 \;\leq\; 2\mu A.$$

The rest of the proof is almost identical, using R modified as explained above and some other minor changes where ρ_k-terms need to be considered, e.g., the term $+\rho_{m_\ell}$ is introduced in (21), yielding an additional sum in (22), which remains finite when passing to the limit because (ρ_k) is summable. □

Similar extensions are possible when using dynamic step sizes of the form (2). The upper bounds (11) and (13) for the projection accuracies (ε_k) will depend on (ρ_k) as well, which of course must be taken into account when extending the proofs accordingly. Then, summability of (ρ_k) (implying $\rho_k \to 0$) is enough to guarantee convergence. In particular, one can again choose $\rho_k \leq \mu\, \varepsilon_k$ for some $\mu > 0$. We will not go into detail here, since the extensions are straightforward.

4.2 Variable target values

From a practical viewpoint, it may be desirable to have an algorithm, using dynamic step sizes, that does not require the user to *know* a priori whether an estimate φ is larger or smaller than f^*, respectively. Moreover, relying on a constant estimate may lead to overly small or large steps, which slows down the convergence process (and, w.r.t. ISA (Algorithm 2), can also lead to many projection accuracy refinement phases). Thus, a typical approach is to replace the constant estimate φ by *variable target values* φ_k. These target values are then updated in the course of the algorithm to increasingly better estimates of f^*, so that the dynamic step size (2) more and more resembles the "ideal" Polyak step size (which would use $\varphi = f^*$). In principle, such extensions are also possible for the ISA framework. We briefly describe the most important aspects in the following.

First, note that Theorems 2 and 3 provide bounds on the projection accuracies (ε_k) needed for convergence; clearly, if it is unknown whether $\varphi_k \geq f^*$ or $\varphi_k < f^*$, one must therefore choose $0 \leq \varepsilon_k \leq \min\{\bar{\varepsilon}_k, |\tilde{\varepsilon}_k|, \nu_k\}$, with $\bar{\varepsilon}_k$ and $\tilde{\varepsilon}_k$ given by (11) and (13), respectively.

Crucial for any variable target value method is the ability to somehow recognize whether $\varphi_k \geq f^*$ or $\varphi_k < f^*$. If all iterates are feasible, this amounts to recognizing whether $X \cap \{x \mid f(x) \leq \varphi_k\} \neq \emptyset$ (or, as $x \in X$, simply $f(x) \leq \varphi_k$), implying $\varphi_k \geq f^*$, or $X \cap \{x \mid f(x) \leq \varphi_k\} = \emptyset$, to infer that $\varphi_k < f^*$, see, e.g., [14]. However, in the case of (possibly) infeasible iterates, $f_k \leq \varphi_k$ does not necessarily imply that φ_k is too large. On the other hand, viewing the ISA iterates x^k as points of the "blown-up" feasible set $\mathcal{B}_X^{\varepsilon_{k-1}} := \{x \mid x = y+z, y \in X, \|z\| \leq \varepsilon_{k-1}\}$, then $\mathcal{B}_X^{\varepsilon_k} \cap \{x \mid f(x) \leq \varphi_k\} = \emptyset$ also implies that $\varphi_k < f^*$, since $X \subseteq \mathcal{B}_X^{\varepsilon_k}$.

In view of Theorem 3, keeping φ_k constant once we recognized that $\varphi_k < f^*$ ensures convergence of (f_k) to f^* (in practice, it may nevertheless be desirable to further improve the estimate φ_k in order to avoid overly large steps in the vicinity of the optimum). The associated case $\mathcal{B}_X^{\varepsilon_k} \cap \{x \mid f(x) \leq \varphi_k\} = \emptyset$ can be detected in practice, see [14, Section III.C] for details in the case of a feasible method; these results are extensible to the ISA framework with appropriate modifications.

The other case, $\varphi_k \geq f^*$, could be detected, e.g., with the help of an estimate of the Lipschitz constant of f (recall that every convex function is locally Lipschitz and useful estimates should usually be available) and the distances to X implied by the projection accuracies.

In the literature, various schemes have been considered as update rules for variable targets (φ_k), see, e.g., [5, 28, 19, 48, 40, 14, 31, 36]. In principle, many such rules could be straightforwardly used in, or adapted to, a variable target value ISA.

4.3 Computable bounds on $d_{X^*}(x^k)$

The results in Theorems 2 and 3 hinge on bounds $\bar{\varepsilon}_k$ and $\tilde{\varepsilon}_k$ on the projection accuracy parameters ε_k, respectively. These bounds depend on unknown information and therefore seem of little practical use such as, for instance, an automated accuracy control in an implementation of the dynamic step size ISA. While the quantity f^* can sometimes be replaced by estimates directly, it will generally be hard to obtain useful estimates for the distance of the current iterate to the optimal set. However, such estimates are available for certain classes of objective functions. We will sketch several examples in the following.

For instance, when f is a *strongly convex function*, i.e., there exists some constant $C > 0$ such that for all x, y and $\mu \in [0, 1]$

$$f(\mu x + (1-\mu)y) \;\le\; \mu f(x) + (1-\mu)f(y) - C\,\mu(1-\mu)\|x-y\|^2,$$

one can use the following upper bound on the distance to the optimal set [28]:

$$d_{X^*}(x) \;\le\; \min\left\{ \sqrt{\tfrac{f(x)-f^*}{C}},\; \tfrac{1}{2C} \min_{h \in \partial f(x)} \|h\| \right\}.$$

For functions f such that $f(x) \ge C\,\|x\| - D$, with constants $C, D > 0$, one can make use of $d_{X^*}(x) \le \|x\| + \frac{1}{C}(f^* + D)$, obtained by simply employing the triangle inequality. Another related example class is induced by coercive self-adjoint operators F, i.e., $f(x) \coloneqq \langle Fx, x\rangle \ge C\|x\|^2$ with some constant $C > 0$ and a scalar product $\langle\cdot,\cdot\rangle$. The (usually) unknown f^* appearing above may again be treated using estimates.

Yet another important class is comprised of functions which have a set of weak sharp minima [18] over X, i.e., there exists a constant $\mu > 0$ such that

$$f(x) - f^* \;\ge\; \mu\, d_{X^*}(x) \qquad \forall\, x \in X. \tag{36}$$

Using $d_{X^*}(x) \le d_X(x) + d_{X^*}(\mathcal{P}_X(x))$ for $x \in \mathbb{R}^n$, we can then estimate the distance of x to X^* via the weak sharp minima property of f. An important subclass of such functions is composed of the polyhedral functions, i.e., f has the form $f(x) = \max\{\, a_i^\top x + b_i \mid 1 \le i \le N \,\}$, where $a_i \ne 0$ for all i; the scalar μ is then given by $\mu = \min\{\, \|a_i\| \mid 1 \le i \le N \,\}$. Rephrasing (36) as

$$d_{X^*}(x) \;\le\; \frac{f(x) - f^*}{\mu} \qquad \forall x \in X,$$

we see that for $\varphi \le f^*$ (e.g., dual lower bounds φ),

$$d_{X^*}(x) \;\le\; \frac{f(x) - \varphi}{\mu} \qquad \forall x \in X.$$

Thus, when the bounds on the distance to the optimal set derived from using the above inequalities become too conservative (i.e., too large, resulting in very small $\tilde{\varepsilon}_k$-bounds), one could try to improve the above bounds by improving the lower bound φ.

In practice on might have access to (problem-specific) estimates of $d_{X^*}(x)$; in [14], it is claimed that “for most problems” prior experience or heuristical considerations can be used to that end. For instance, if X is compact, the diameter of X leads to the (conservative) estimate $d_{X^*}(x) \le \operatorname{diam}(X) + d_X(x)$.

5 Examples

In this section, we briefly discuss two examples in which we can design adaptive approximate projections as considered in the ISA framework. In the first example, we focus on the theoretical aspects of how our notion of adaptive approximate projection could be used to handle a certain class of constraints appearing in stochastic programs. The second application considers a (deterministic) optimization problem for which we specialize ISA and present some numerical experiments.

5.1 Convex expected value constraints

We consider *expected value constraints* [47, 33] of the following form

$$g(x) \coloneqq \mathbb{E}[f(x;\omega)] = \int_{\Omega} f(x;\omega)\, p(\omega)\, d\omega \leq \eta, \tag{37}$$

where $\mathbb{E}$ denotes the expected value, $\omega \in \Omega \subseteq \mathbb{R}^q$ is a vector of random variables with density p, x are deterministic variables in $\mathbb{R}^n$, $f : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}$, and $\eta \in \mathbb{R}$. If f is convex in x for every $\omega \in \Omega$, (37) is a convex constraint. Expected value constraint appear in stochastic programming as, for instance, the expectational form of chance constraints, see, e.g., [11, 7], or when modeling expected loss or Value-at-Risk via integrated chance constraints, see, e.g., [21, 27, 22].

While generally $g(x)$ cannot be easily computed exactly, it can be approximated using Monte Carlo methods, if samples of ω can be (cheaply) generated. Here, taking M independent samples $\omega^1, \dots, \omega^M$, yields the approximation

$$\hat{g}_M(x) \coloneqq \frac{1}{M} \sum_{i=1}^{M} f(x;\omega^i) \tag{38}$$

of $g(x)$. Moreover, we assume that we can compute a subgradient $G(x;\omega) \in \partial_x f(x;\omega)$ for each value of x and ω. Thus, we have $h \coloneqq \mathbb{E}[G(x;\omega)] \in \partial g(x)$. We then use the approximation

$$\hat{h}_M(x) \coloneqq \frac{1}{M} \sum_{i=1}^{M} G(x;\omega^i), \tag{39}$$

which is a "noisy unbiased subgradient" of g at x; see [8] for details.

Considering the Lagrangean $L(y,\lambda) = \frac{1}{2}\|x-y\|^2 + \lambda\,(g(y)-\eta)$ of the projection problem for some point x and the set of feasible points w.r.t. (37), the optimality conditions for the projection obtained by differentiating L are

$$-x + y + \lambda\, h = 0, \quad \text{for some } h \in \partial g(y), \tag{40}$$

$$g(y) - \eta = 0. \tag{41}$$

Then, the idea is to replace $g(y)$ and h by the estimates $\hat{g}_M(y)$ and $\hat{h}_M(y)$, respectively. An adaptive approximate projection is obtained by solving

$$y = x - \lambda\,\hat{h}_M(y), \qquad \hat{g}_M(y) = \eta. \tag{42}$$

For an appropriate sampling process, we can adaptively keep control on the resulting projection error (with high probability).

We now demonstrate this approach on a simple example constraint in which the above system can be solved easily and we obtain explicit projection error bounds: Consider a linear function with random coefficients, i.e., $f(x;\omega) = \omega^\top x$ and $q = n$. This particular type of constraint is closely related to integrated chance constraints which are used, for instance, to model bounds on expected losses of some sort; see, e.g., [21, 27]. For this choice of f, our Monte Carlo estimates are

$$\hat{h}_M(x) = \hat{h}_M = \frac{1}{M}\sum_{i=1}^{M}\omega^i \qquad \text{and} \qquad \hat{g}_M(x) = \hat{h}_M^\top x. \tag{43}$$

Note that if $\mathbb{E}[\hat{h}_M(x)]$ is unknown, the feasibility operator construction in [23] is not applicable. Moreover, assuming h, $\hat{h}_M \neq 0$ corresponds to imposing a lower bound on the subgradient norm, like in the convergence theorems for ISA. Observing that $\hat{h}_M$ is independent of x (so in particular, $\hat{h}_M(y) = \hat{h}_M$ as well), we can solve (42) to obtain the solution

$$\mathcal{P}^M(x) \coloneqq x - \left(\frac{\hat{h}_M^\top x - \eta}{\|\hat{h}_M\|^2}\right)\hat{h}_M \tag{44}$$

to the approximated projection problem. The exact projection is given by

$$\mathcal{P}^\infty(x) \coloneqq x - \frac{h^\top x - \eta}{\|h\|^2}h, \tag{45}$$

and—as the notation suggests—we have $\mathcal{P}^\infty(x) = \lim_{M\to\infty}\mathcal{P}^M(x)$ almost-surely, since $\text{Prob}(\lim_{M\to\infty}\hat{h}_M = h) = 1$ by the (strong) law of large numbers.

For sufficiently large M, we can use explicit $(1-\alpha)$-confidence intervals for the expected value $h = \mathbb{E}[\hat{h}_M]$ via the central limit theorem, and eventually obtain

$$\text{Prob}\big(\ \|\mathcal{P}^M(x) - \mathcal{P}^\infty(x)\| \le \varepsilon_M\ \big) = 1 - \alpha, \tag{46}$$

where

$$\varepsilon_M \coloneqq \left\|\frac{\hat{h}_M^\top x - \eta}{\|\hat{h}_M\|^2}\hat{h}_M - \frac{\hat{h}_M^\top x - \eta + \overline{c}\cdot q_M^\top x}{\|\hat{h}_M + \overline{c}\cdot q_M\|^2}(\hat{h}_M + \overline{c}\cdot q_M)\right\|,$$

with $\overline{c} = -\text{sign}(\hat{h}_M^\top q_M)$ and

$$q_M = \frac{q_{(1-\alpha/2)}}{\sqrt{M}\sqrt{M-1}}\left(\sqrt{\sum_{i=1}^{M}((\omega^i)_1 - (\hat{h}_M)_1)^2}, \ldots, \sqrt{\sum_{i=1}^{M}((\omega^i)_n - (\hat{h}_M)_n)^2}\right)^\top,$$

where $q_{(1-\alpha/2)}$ denotes the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. Thus, for any given $\alpha \in (0,1)$ and for sufficiently large M, $\mathcal{P}^M$ defines an adaptive approximate projection operator as specified in the ISA framework, with probability $1-\alpha$.

It is noteworthy that the projection accuracy directly depends on M, and in the linear example above we could iteratively refine the estimate $\hat{h}_M$ easily by incorporating newly drawn independent samples.

5.2 Compressed sensing

Compressed Sensing (CS) is a recent and very active research field dealing, loosely speaking, with the recovery of signals from incomplete measurements. We refer the interested reader to [17, 9, 15] for more information, surveys, and key literature. A core problem of CS is finding the sparsest solution to an underdetermined linear system, i.e.,

$$\min \|x\|_0 \quad \text{s.t.} \quad Ax = b, \qquad (A \in \mathbb{R}^{m\times n},\ \operatorname{rank}(A) = m,\ m < n), \tag{47}$$

where $\|x\|_0$ denotes the ℓ_0 quasi-norm or support size of the vector x, i.e., the number of its nonzero entries. This problem is known to be $\mathcal{NP}$-hard. Hence, a common approach is considering the convex relaxation known as ℓ_1-minimization or Basis Pursuit [12]:

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = b. \tag{48}$$

It was shown that under certain conditions, the solutions of (48) and (47) coincide, see, e.g., [10, 17]. This motivated a large amount of research on the efficient solution of (48), especially in large-scale settings. In this section, we briefly outline a specialization of the ISA to the ℓ_1-minimization problem (48) and present some numerical experiments indicating that the algorithm is an interesting candidate in the context of Compressed Sensing.

Subgradients. The subdifferential of the ℓ_1-norm at a point x is given by

$$\partial\|x\|_1 = \left\{ h \in [-\mathbb{1},\mathbb{1}]^n \;\middle|\; h_i = \frac{x_i}{|x_i|}, \quad \forall\, i \in \{1,\dots,n\} \text{ with } x_i \neq 0 \right\}. \tag{49}$$

We may therefore simply use the signs of the iterates as subgradients, i.e.,

$$\partial\|x^k\|_1 \ni h^k \;:=\; \operatorname{sign}(x^k) \;=\; \begin{cases} 1, & (x^k)_i > 0, \\ 0, & (x^k)_i = 0, \\ -1, & (x^k)_i < 0. \end{cases} \tag{50}$$

As long as $b \neq 0$, the upper and lower bounds on the norms of the subgradients satisfy $\underline{H} \geq 1$ and $\overline{H} \leq n$.

Adaptive approximate projection. For linear equality constraints as in (48), the Euclidean projection of a point $z \in \mathbb{R}^n$ onto the affine feasible set $X \coloneqq \{\, x \mid Ax = b \,\}$ can be explicitly calculated as

$$\mathcal{P}_X(z) \;=\; \big(I - A^\top(AA^\top)^{-1}A\big)z + A^\top(AA^\top)^{-1}\,b, \tag{51}$$

where I denotes the $(n \times n)$ identity matrix. However, for numerical stability, we wish to avoid the explicit calculation of the projection matrix because it involves determining the inverse of the matrix product $AA^\top$. Instead of applying (51) in each iteration, we can use the following adaptive procedure:

$$z^k \coloneqq x^k - \alpha_k h^k \qquad \text{(unprojected next iterate)}, \tag{52}$$

$$\text{find an approximate solution } q^k \text{ of } AA^\top q = Az^k - b, \tag{53}$$

$$x^{k+1} \coloneqq z^k - A^\top q^k. \tag{54}$$

Note that the matrix $AA^\top$ is symmetric and positive definite, for A with full (row-)rank m. Hence, the linear system in (53) can be solved by an iterative method, e.g., the method of Conjugate Gradients (CG) [24].

For a given ε_k, stopping the CG procedure in (53) as soon as the iteratively updated approximate solution q^k satisfies

$$\|AA^\top q^k - \big(A(x^k - \alpha_k h^k) - b\big)\|_2 \;\le\; \sigma_{\min}(A)\,\varepsilon_k, \tag{55}$$

where $\sigma_{\min}(A) > 0$ is the smallest singular value of A, ensures that (52)–(54) form an adaptive approximate projection operator of the type (7). Note that a truncated CG procedure (with any fixed number of iterations) can also be shown to define a "feasibility operator" of the type considered in [23].

Furthermore, to obtain computable upper bounds on (ε_k), we can use the results about weak sharp minima discussed in the previous section: The ℓ_1-norm can be rewritten as a polyhedral function. With $\varphi \le f^*$ (which is easily available, e.g., $\varphi = 0$), we can thus derive

$$d_{X^*}(x^k) \;\le\; 2\frac{\|Ax^k - b\|_2}{\sigma_{\min}(A)} + \frac{\|x^k\|_1 - \varphi}{\sqrt{n}}.$$

In total, this yields bounds that can be easily computed from the original data only. Theorems 1, 2, or 3 then provide explicit convergence statements.

Numerical Experiments It is well-known that (48) can be solved as a linear program (LP), e.g., employing the standard variable split $x = x^+ - x^-$:

$$\min\, x^+ + x^- \quad \text{s.t.} \quad Ax^+ - Ax^- = b,\; x^+ \ge 0,\; x^- \ge 0. \tag{56}$$

Another common approach to (48) is to solve a sequence of regularized problems of the form

$$\min\, \tfrac{1}{2}\|Ax - b\|_2^2 + \tau\|x\|_1, \tag{57}$$

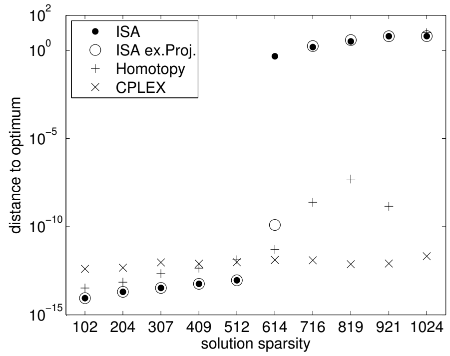
with decreasing τ. As $\tau \to 0$, the solution sequence $x(\tau)$ of (57) converges to a solution of (48). The homotopy method (see, e.g., [43, 39]) traces this solution path for decreasing τ and has the desirable property to require only k steps to reach the optimal solution x^* to (48), if x^* has only k nonzero entries and k is sufficiently small.

We performed experiments to compare our ISA Algorithm 2, applied to (48) (using adaptive approximate or exact projections), with the commercial LP-solver Cplex 12.5 (dual simplex method applied to (56)) and the homotopy implementation (version 1.0) available at `http://users.ece.gatech.edu/~sasif/homotopy/`. In our ISA implementation we employ at most 5 CG steps to approximate the projection; albeit differing from theory, this turned out to suffice. Moreover, the subgradients are stabilized as in [37], and the parameter λ_k is halved after 5 consecutive iterations without relevant improvement of the objective ($\lambda_0 = 0.85$); the method terminates when the step sizes become too small or if a stagnation of the algorithmic process is detected. By stagnation, we mean that either the objective improvement stalls over a span of 500 iterations, or the approximate support $S = \{i : |x_i^k| > \max\{10^{-6}, s\}\}$ does not change over 10 successive updates, which are performed every $m/100$ iterations; here s is chosen such that the entries x_j^k with $|x_j^k| \geq s$ account for at least 99.99% of $\|x^k\|_1$. Finally, as a postprocessing step after termination, we try to improve the solution by solving the system restricted to columns indexed by S, similar to the "debiasing" step described in [51, Section II.I].

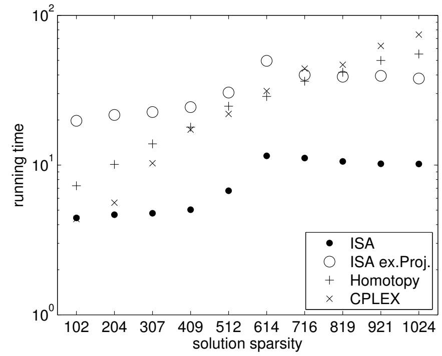
Note that in contrast to Cplex, the homotopy method and ISA are implemented in Matlab (version R2012a/7.14). Moreover, by default, Cplex ensures feasibility in the sense that the computed solution $\overline{x}$ obeys $\|A\overline{x} - b\|_\infty \leq 10^{-6}$; from the respective convergence results, both the homotopy method and ISA will reach this level of feasibility after finitely many iterations. As a safeguard, we added an additional high-accuracy projection after regular termination. However, this step was not required for the homotopy method, and only on a single instance for ISA (this induced additional running time and the time for the postprocessing step is incorporated in the times reported below).

The first test uses a 1024×4096 Gaussian matrix, the second one a partial discrete cosine transform (DCT) matrix consisting of 512 randomly drawn rows of the 2048×2048 DCT matrix; all columns are normalized to unit Euclidean length. For both matrices, we constructed ten vectors x^i with sparsities $\|x^i\|_0 = i \cdot m/10$, $i \in \{1, \dots, 10\}$, (rounded down to the next integer value). The nonzero entries are ± 1 and each x^i is the known unique solution to the instance given by the respective matrix A and right hand side vector $b \coloneqq Ax^i$, where uniqueness was achieved by ensuring the "strong source condition" (see, e.g., [20]) by means of the methodology proposed in [32].

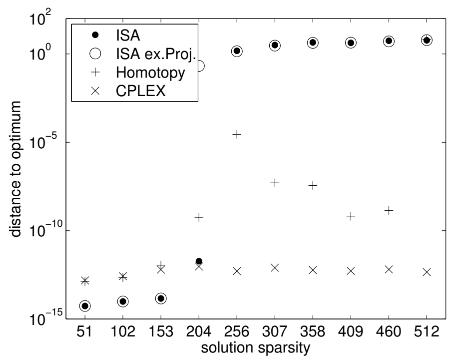
Figure 3 shows the running times (in seconds) and the ℓ_2-norm distances to the respective known optimal solution. As explained above, all solutions are feasible to within an ℓ_∞-tolerance of 10^{-6}. The experiments show that using adaptive approximate projections instead of the exact ones in ISA saves a considerable amount of time, as was to be expected. The achieved final accuracy



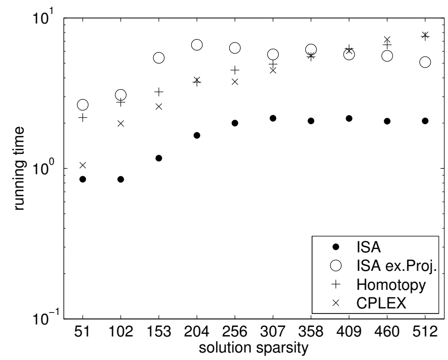
(a) ℓ_2-distances to optimum for instances with 1024×4096 Gaussian matrix.



(b) Running times (s) for instances with 1024×4096 Gaussian matrix.



(c) ℓ_2-distances to optimum for instances with 512×2048 partial DCT matrix.



(d) Running times (s) for instances with 512×2048 partial DCT matrix.

Fig. 3. Numerical experiments for Gaussian matrix ((a) and (b)) and partial DCT matrix ((c) and (d)), each with normalized columns, for varying solution sparsities.

is almost always (nearly) the same. For the varying sparsity levels of the solution, we see that all solvers struggle when the number of nonzero entries in the optimum exceeds about $m/2$: Cplex and the homotopy method still produce mostly accurate solutions but at the cost of a significant increase in the required solution times (note the logarithmic scales on the vertical axes), ISA on the other hand has a somewhat more stable runtime behavior, but loses accuracy when the solution is dense.

Since in Compressed Sensing, the solutions encountered are typically very sparse, the interesting cases are those with sparsity (much) smaller than $m/2$. Clearly, for such sparse optimal solutions, ISA (with adaptive approximate projections) is superior to Cplex and the homotopy implementation both in terms of accuracy and speed. Thus, these examples show the potential of ISA as a successful algorithm for CS sparse recovery.

6 Concluding remarks

Several aspects remain subject to future research. For instance, it would be interesting to investigate whether our framework extends to (infinite-dimensional) Hilbert space settings, incremental subgradient schemes, bundle methods (see, e.g., [25, 29]), or Nesterov's algorithm [42]. It is also of interest to consider how the ISA framework could be combined with error-admitting settings such as those in [52, 41], i.e., for random or deterministic (non-vanishing) noise and erroneous function or subgradient evaluations. Some of the recent results in [41], which all require feasible iterates, seem conceptually close to our convergence analyses, so we presume a blend of the two approaches to be rather fruitful. It would also be of interest to investigate convergence behavior with other general notions of "adaptive approximate projections", e.g., solving the projection problem with an approximation algorithm with additive or multiplicative performance guarantee.

From a practical viewpoint, it will be interesting to see how ISA, or possibly a variable target value variant as described in Section 4.2, compares with other solvers in terms of solution accuracy and runtime. For the ℓ_1-minimization problem (48), we have seen in Section 5.2 that ISA promises to be an interesting candidate; an extensive computational comparison of various state-of-the-art ℓ_1-solvers, including (a more refined version of) our ISA implementation, can be found in [38]. An extensive test for convex expected value constraints, while beyond the scope of this paper, would be an interesting further line of work.

Acknowledgments. We thank the anonymous referees for their numerous helpful comments which greatly helped improving this paper.

References

1. Y. I. Alber, A. N. Iusem, and M. V. Solodov, *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space*, Mathematical Programming, 81 (1998), pp. 23–35.
2. E. Allen, R. Helgason, J. Kennington, and B. Shetty, *A generalization of Polyak's convergence result for subgradient optimization*, Mathematical Programming, 37 (1987), pp. 309–317.
3. K. M. Anstreicher and L. A. Wolsey, *Two "well-known" properties of subgradient optimization*, Mathematical Programming, 120 (2009), pp. 213–220.
4. H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review, 38 (1996), pp. 367–426.
5. M. S. Bazaraa and H. D. Sherali, *On the choice of step size in subgradient optimization*, European Journal of Operations Research, 7 (1981), pp. 380–388.
6. D. P. Bertsekas and S. K. Mitter, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, SIAM Journal of Control, 11 (1973), pp. 637–652.
7. J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer Series in Operations Research, Springer, 1999. corrected second printing.
8. S. Boyd and A. Mutapcic, *Stochastic subgradient methods.* Lecture notes, `http://see.stanford.edu/materials/lsocoee364b/04-stoch_subgrad_notes.pdf`, 2007.

9. A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Review, 51 (2009), pp. 34–81.
10. E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory, 52 (2006), pp. 489–509.
11. A. Charnes and W. W. Cooper, *Chance-constrained programming*, Management Science, 6 (1959), pp. 73–79.
12. S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.
13. A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods. II. Beyond the elliptic case*, Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics, 2 (2002), pp. 203–245.
14. P. L. Combettes and J. Luo, *An adaptive level set method for nondifferentiable constrained image recovery*, IEEE Transactions on Image Processing, 11 (2002), pp. 1295–1304.
15. *Compressive sensing resources.* `http://dsp.rice.edu/cs` (Accessed 03/26/2012).
16. G. D'Antonio and A. Frangioni, *Convergence analysis of deflected conditional approximate subgradient methods*, SIAM Journal on Optimization, 20 (2009), pp. 357–386.
17. D. L. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306.
18. M. C. Ferris, *Weak sharp minima and exact penalty functions*, Tech. Rep. 779, Comp. Sci. Dept., University of Wisconsin, Madison, WI, 1988.
19. J. L. Goffin and K. Kiwiel, *Convergence of a simple subgradient level method*, Mathematical Programming, 85 (1999), pp. 207–211.
20. M. Grasmair, M. Haltmeier, and O. Scherzer, *Necessary and sufficient conditions for linear convergence of ℓ^1-regularization*, Communications on Pure and Applied Mathematics, 64 (2011), pp. 161–182.
21. W. K. K. Haneveld, *Duality in Stochastic Linear and Dynamic Programming*, vol. 274 of Lecture Notes in Economics and Mathematical Systems, Springer, 1986.
22. W. K. K. Haneveld and M. H. van der Vlerk, *Integrated chance constraints: reduced forms and an algorithm*, Computational Management Science, 3 (2006), pp. 245–269.
23. E. S. Helou Neto and A. R. De Pierro, *Incremental subgradients for constrainted convex optimization: A unified framework and new methods*, SIAM Journal on Optimization, 20 (2009), pp. 1547–1572.
24. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 409–436.
25. J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms. II*, vol. 306 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer, 1993.
26. ———, *Fundamentals of Convex Analysis*, Springer, 2004. Corrected Second Printing.
27. P. Kall and J. Mayer, *Stochastic Linear Programming. Models, Theory, and Computation*, Springer, 2005.
28. S. Kim, H. Ahn, and S.-C. Cho, *Variable target value subgradient method*, Mathematical Programming, 49 (1991), pp. 359–369.
29. K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Mathematical Programming, 46 (1990), pp. 105–122.

30. ———, *Subgradient method with entropic projections for convex nondifferentiable minimization*, Journal on Optimization Theory and Applications, 96 (1998), pp. 159–173.
31. ———, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM Journal on Optimization, 14 (2004), pp. 807–840.
32. C. Kruschel and D. A. Lorenz, *Maximal recoverable supports for sparse recovery.* Preprint, 2013.
33. D. Kuhn, *Convergent bounds for stochastic programs with expected value constraints*, Journal of Optimization Theory and Applications, 141 (2009), pp. 597–618.
34. T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization – theory and applications*, European Journal of Operations Research, 88 (1996), pp. 382–403.
35. A. S. Lewis, D. R. Luke, and J. Malick, *Local linear convergence for alternating and averaged nonconvex projections*, Foundations of Computational Mathematics, 9 (2009), pp. 485–513.
36. C. Lim and H. D. Sherali, *Convergence and computational analyses for some variable target value and subgradient deflection methods*, Computational Optimization and Applications, 34 (2005), pp. 409–428.
37. A. Löbel, *Optimal Vehicle Scheduling in Public Transit*, PhD thesis, Technische Universität Berlin, 1998. Shaker.
38. D. A. Lorenz, M. E. Pfetsch, and A. M. Tillmann, *Solving Basis Pursuit: Subgradient algorithm, heuristic optimality check, and solver comparison.* Optimization Online E-Print ID 2011-07-3100, 2011.
39. D. Malioutov, M. Çetin, and A. Willsky, *Homotopy continuation for sparse signal representation*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), vol. 5, 2005, pp. 733–736.
40. A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization, 12 (2001), pp. 109–138.
41. ———, *The effect of deterministic noise in subgradient methods*, Mathematical Programming, 125 (2010), pp. 75–99.
42. Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.
43. M. Osbourne, B. Presnell, and B. Turlach, *A new approach to variable selection in least squares problems*, IMA Journal of Numerical Analysis, 20 (2000), pp. 389–402.
44. B. T. Polyak, *A general method for solving extremal problems*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 33–36.
45. ———, *Minimization of nonsmooth functionals*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.
46. ———, *Subgradient methods: A survey of soviet research*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., IIASA Proceedings Series, Pergamon Press, 1978, pp. 5–29.
47. A. Prékopa, *Contributions to the theory of stochastic programming*, Mathematical Programming, 4 (1973), pp. 202–221.
48. H. D. Sherali, G. Choi, and C. H. Tubcbilek, *A variable target value method for nondifferentiable optimization*, Operations Research Letters, 26 (2000), pp. 1–8.
49. N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer, 1985.

50. E. van den Berg, M. Schmidt, M. P. Friedlander, and K. Murphy, *Group sparsity via linear-time projection*, Tech. Rep. TR-2008-09, University of British Columbia, 2008.
51. S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by seperable approximation*, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493.
52. A. J. Zaslavski, *The projected subgradient method for nonsmooth convex optimization in the presence of computational error*, Numerical Functional Analysis and Optimization, 31 (2010), pp. 616–633.