

Incremental proximal methods for large scale convex optimization

Dimitri P. Bertsekas

Received: 1 August 2010 / Accepted: 13 March 2011 / Published online: 11 June 2011
© Springer and Mathematical Optimization Society 2011

Abstract We consider the minimization of a sum $\sum_{i=1}^m f_i(x)$ consisting of a large number of convex component functions f_i . For this problem, incremental methods consisting of gradient or subgradient iterations applied to single components have proved very effective. We propose new incremental methods, consisting of proximal iterations applied to single components, as well as combinations of gradient, subgradient, and proximal iterations. We provide a convergence and rate of convergence analysis of a variety of such methods, including some that involve randomization in the selection of components. We also discuss applications in a few contexts, including signal processing and inference/machine learning.

Keywords Proximal algorithm · Incremental method · Gradient method · Convex

Mathematics Subject Classification (2000) 90C33 · 90C90

1 Introduction

In this paper we focus on problems of minimization of a cost consisting of a large number of component functions, such as

Laboratory for Information and Decision Systems Report LIDS-P-2847, August 2010 (revised March 2011); to appear in Math. Programming Journal, 2011. Research supported by AFOSR Grant FA9550-10-1-0412. Many thanks are due to Huizhen (Janey) Yu for extensive helpful discussions and suggestions.

D. P. Bertsekas (✉)
Department of Electrical Engineering and Computer Science,
Laboratory for Information and Decision Systems, M.I.T., Mass,
Cambridge, MA 02139, USA
e-mail: dimitrib@MIT.EDU

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^m f_i(x) \\ & \text{subject to} && x \in X, \end{aligned} \tag{1}$$

where $f_i : \mathfrak{R}^n \mapsto \mathfrak{R}$, $i = 1, \dots, m$, are convex, and X is a closed convex set.¹ When m , the number of component functions, is very large there is an incentive to use incremental methods that operate on a single component f_i at each iteration, rather than on the entire cost function. If each incremental iteration tends to make reasonable progress in some “average” sense, then depending on the value of m , an incremental method may significantly outperform (by orders of magnitude) its nonincremental counterpart, as experience has shown.

Additive cost problems of the form (1) arise in many contexts such as dual optimization of separable problems, machine learning (regularized least squares, maximum likelihood estimation, the EM algorithm, neural network training), and others (e.g., distributed estimation, the Fermat–Weber problem in location theory, etc). They also arise in the minimization of an expected value that depends on x and some random vector; then the sum $\sum_{i=1}^m f_i(x)$ is either an expected value with respect to a discrete distribution, as for example in stochastic programming, or is a finite sample approximation to an expected value. The author’s paper [16] surveys applications that are well-suited for the incremental approach. In the case where the components f_i are differentiable, incremental gradient methods take the form

$$x_{k+1} = P_X(x_k - \alpha_k \nabla f_{i_k}(x_k)), \tag{2}$$

where α_k is a positive stepsize, $P_X(\cdot)$ denotes projection on X , and i_k is the index of the cost component that is iterated on. Such methods have a long history, particularly for the unconstrained case ($X = \mathfrak{R}^n$), starting with the Widrow–Hoff least mean squares (LMS) method [58] for positive semidefinite quadratic component functions (see e.g., [35], and [7, Section 3.2.5]). For nonquadratic cost components, such methods have been used extensively for the training of neural networks under the generic name “backpropagation methods.” There are several variants of these methods, which differ in the stepsize selection scheme, and the order in which components are taken up for iteration (it could be deterministic or randomized). They are related to gradient methods with errors in the calculation of the gradient, and are supported by convergence analyses under various conditions; see Luo [35], Grippo [26, 27], Luo and Tseng [34], Mangasarian and Solodov [36], Bertsekas [12, 13], Solodov [54], Tseng [55]. An alternative method that also computes the gradient incrementally, one component per iteration, is proposed by Blatt et al. [1]. Stochastic versions of these methods also have a long history, and are strongly connected with stochastic approximation methods. The main difference between stochastic and deterministic formulations is that the former involve sampling a sequence of cost components from an infinite population

¹ Throughout the paper, we will operate within the n -dimensional space \mathfrak{R}^n with the standard Euclidean norm, denoted $\|\cdot\|$. All vectors are considered column vectors and a prime denotes transposition, so $x'x = \|x\|^2$. Throughout the paper we will be using standard terminology of convex optimization, as given for example in textbooks such as Rockafellar’s [50], or the author’s recent book [15].

under some statistical assumptions, while in the latter the set of cost components is predetermined and finite.

Incremental gradient methods typically have a slow asymptotic convergence rate not only because they are first order gradient-like methods, but also because they require a diminishing stepsize [such as $\alpha_k = O(1/k)$] for convergence. If α_k is instead taken to be constant, an oscillation whose size depends on α_k typically arises, as shown by [35]. These characteristics are unavoidable in incremental methods, and are typical of all the methods to be discussed in this paper. However, because of their frequently fast initial convergence rate, incremental methods are often favored for large problems where great solution accuracy is not of paramount importance (see [14] and [16] for heuristic arguments supporting this assertion).

Incremental subgradient methods apply to the case where the component functions f_i are convex and nondifferentiable. They are similar to their gradient counterparts (2) except that an arbitrary subgradient $\tilde{\nabla} f_{i_k}(x_k)$ of the cost component f_{i_k} is used in place of the gradient:²

$$x_{k+1} = P_X \left(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k) \right). \tag{3}$$

Such methods were proposed in the Soviet Union by Kibardin [30], following the earlier paper by Litvakov [33] (which considered convex nondifferentiable extensions of linear least squares problems) and other related subsequent proposals. These works remained unnoticed in the Western literature, where incremental methods were reinvented often in different contexts and with different lines of analysis; see Solodov and Zavriev [53], Ben-Tal et al. [4], Nedić and Bertsekas [39–41], Nedić et al. [38], Kiwiel [31], Rabbat and Nowak [48,49], Shalev-Shwartz et al. [52], Johansson et al. [29], Helou and De Pierro [28], Predd et al. [44], and Ram et al. [46,47]. Incremental subgradient methods have convergence properties that are similar in many ways to their gradient counterparts, the most important similarity being the necessity of a diminishing stepsize α_k for convergence. The lines of analysis, however, tend to be different, since incremental gradient methods rely for convergence on the decrease of the cost function value, while incremental gradient methods rely on the decrease of the iterates’ distance to the optimal solution set. The line of analysis of the present paper is of the latter type, similar to earlier works of the author and his collaborators (see [38–40], and the textbook presentation in [5]).

In this paper, we propose and analyze for the first time incremental methods that relate to proximal algorithms. The simplest one for problem (1) is of the form

$$x_{k+1} = \arg \min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}, \tag{4}$$

which bears the same relation to the proximal minimization algorithm (Martinet [37], Rockafellar [51]) as the incremental subgradient method (3) bears to the classical

² In this paper, we use $\tilde{\nabla} f(x)$ to denote a subgradient of a convex function f at a vector x . The choice of $\tilde{\nabla} f(x)$ from within $\partial f(x)$ is clear from the context.

subgradient method.³ Here $\{\alpha_k\}$ is a positive scalar sequence, and we assume that each $f_i : \mathfrak{R}^n \mapsto \mathfrak{R}$ is a convex function and X is a closed convex set. The motivation for this method is that with a favorable structure of the components, the proximal iteration (3) may be obtained in closed form or be relatively simple, in which case it may be preferable to a gradient or subgradient iteration. In this connection, we note that generally, proximal iterations are considered more stable than gradient iterations; for example in the nonincremental case, they converge essentially for any choice of α_k , but this is not so for gradient methods.

While some cost function components may be well suited for a proximal iteration, others may not be, so it makes sense to consider combinations of gradient/subgradient and proximal iterations. In fact nonincremental combinations of gradient and proximal methods for minimizing the sum of two functions f and h (or more generally, finding a zero of the sum of two nonlinear operators) have a long history, dating to the splitting algorithms of Lions and Mercier [32], and Passty [45], and have become popular recently (see Beck and Teboulle [9, 10], and the references they give to specialized algorithms, such as shrinkage/thresholding, cf. Sect. 5.1).

In this paper we adopt a unified analytical framework that includes incremental gradient, subgradient, and proximal methods, and their combinations, and highlights their common behavior. In particular, we consider the problem

$$\begin{aligned} \text{minimize} \quad & F(x) \stackrel{\text{def}}{=} \sum_{i=1}^m F_i(x) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{5}$$

where for all i , F_i is of the form

$$F_i(x) = f_i(x) + h_i(x), \tag{6}$$

$f_i : \mathfrak{R}^n \mapsto \mathfrak{R}$ and $h_i : \mathfrak{R}^n \mapsto \mathfrak{R}$ are real-valued convex (and hence continuous) functions, and X is a nonempty closed convex set. We implicitly assume here that the functions f_i are suitable for a proximal iteration, while the functions h_i are not and thus may be preferably treated with a subgradient iteration.

One of our algorithms has the form

$$\begin{aligned} z_k &= \arg \min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}, \\ x_{k+1} &= P_X \left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) \right), \end{aligned} \tag{7}$$

where $\tilde{\nabla} h_{i_k}(z_k)$ is an arbitrary subgradient of h_{i_k} at z_k . Note that the iteration is well-defined because the minimum in Eq. (7) is uniquely attained since f_i is continuous

³ In this paper we restrict attention to proximal methods with the quadratic regularization term $\|x - x_k\|^2$. Our approach is applicable in principle when a nonquadratic term is used instead in order to match the structure of the given problem. The discussion of such alternative algorithms is beyond our scope, but the analysis of this paper may serve as a guide for their investigation.

and $\|x - x_k\|^2$ is real-valued, strictly convex, and coercive, while the subdifferential $\partial h_i(z_k)$ is nonempty since h_i is real-valued. Note also that by choosing all the f_i or all the h_i to be identically zero, we obtain as special cases the subgradient and proximal iterations (3) and (4), respectively.

Both iterations (7) and (8) maintain the sequences $\{z_k\}$ and $\{x_k\}$ within the constraint set X , but it may be convenient to relax this constraint for either the proximal or the subgradient iteration, thereby requiring a potentially simpler computation. This leads to the algorithm

$$z_k = \arg \min_{x \in \mathbb{R}^n} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}, \tag{9}$$

$$x_{k+1} = P_X \left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) \right), \tag{10}$$

where the restriction $x \in X$ has been omitted from the proximal iteration, and the algorithm

$$z_k = x_k - \alpha_k \tilde{\nabla} h_{i_k}(x_k), \tag{11}$$

$$x_{k+1} = \arg \min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - z_k\|^2 \right\}, \tag{12}$$

where the projection onto X has been omitted from the subgradient iteration. It is also possible to use different stepsize sequences in the proximal and subgradient iterations, but for notational simplicity we will not discuss this type of algorithm. Still another possibility is to replace h_{i_k} by a linear approximation in an incremental proximal iteration that also involves f_{i_k} . This leads to the algorithm

$$x_{k+1} = \arg \min_{x \in X} \left\{ f_{i_k}(x) + h_{i_k}(x_k) + \tilde{\nabla} h_{i_k}(x_k)'(x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}, \tag{13}$$

which also yields as special cases the subgradient and proximal iterations (3) and (4), when all the f_i or all the h_i are identically zero, respectively.

All of the incremental proximal algorithms given above are new to our knowledge. The closest connection to the existing proximal methods literature is the differentiable nonincremental case of the algorithm (13) (h_i is differentiable, possibly nonconvex, with Lipschitz continuous gradient, and $m = 1$), which has been called the “proximal gradient” method, and has been analyzed and discussed recently in the context of several machine learning applications by Beck and Teboulle [9, 10] (it can also be interpreted in terms of splitting algorithms [32, 45]). We note that contrary to subgradient and incremental methods, the proximal gradient method does not require a diminishing stepsize for convergence to the optimum. In fact, the line of convergence analysis of Beck and Teboulle relies on the differentiability of h_i and the nonincremental character of the proximal gradient method, and is thus different from ours.

Aside from the introduction of a unified incremental framework within which proximal and subgradient methods can be embedded and combined, the purpose of the paper is to establish the convergence properties of the incremental methods (7)–(8),

(9)–(10), (11)–(12), and (13). This includes convergence within a certain error bound for a constant stepsize, exact convergence to an optimal solution for an appropriately diminishing stepsize, and improved convergence rate/iteration complexity when randomization is used to select the cost component for iteration. In Sect. 2, we show that proximal iterations bear a close relation to subgradient iterations, and we use this relation to write our methods in a form that is convenient for the convergence analysis. In Sect. 3 we discuss convergence with a cyclic rule for component selection. In Sect. 4, we discuss a randomized component selection rule and we demonstrate a more favorable convergence rate estimate over the cyclic rule, as well as over the classical nonincremental subgradient method. In Sect. 5 we discuss some applications.

2 Incremental subgradient-proximal methods

We first show some preliminary properties of the proximal iteration in the following proposition. These properties have been commonly used in the literature, but for convenience of exposition, we collect them here in the form we need them. Part (a) provides a key fact about incremental proximal iterations. It shows that they are closely related to incremental subgradient iterations, with the only difference being that the subgradient is evaluated at the end point of the iteration rather than at the start point. Part (b) of the proposition provides an inequality that is useful for our convergence analysis. In the following, we denote by $\text{ri}(S)$ the relative interior of a convex set S , and by $\text{dom}(f)$ the effective domain $\{x \mid f(x) < \infty\}$ of a function $f : \mathfrak{R}^n \mapsto (-\infty, \infty]$.

Proposition 1 *Let X be a nonempty closed convex set, and let $f : \mathfrak{R}^n \mapsto (-\infty, \infty]$ be a closed proper convex function such that $\text{ri}(X) \cap \text{ri}(\text{dom}(f)) \neq \emptyset$. For any $x_k \in \mathfrak{R}^n$ and $\alpha_k > 0$, consider the proximal iteration*

$$x_{k+1} = \arg \min_{x \in X} \left\{ f(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}. \tag{14}$$

(a) *The iteration can be written as*

$$x_{k+1} = P_X \left(x_k - \alpha_k \tilde{\nabla} f(x_{k+1}) \right), \quad i = 1, \dots, m, \tag{15}$$

where $\tilde{\nabla} f(x_{k+1})$ is some subgradient of f at x_{k+1} .

(b) *For all $y \in X$, we have*

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (f(x_{k+1}) - f(y)) - \|x_k - x_{k+1}\|^2 \\ &\leq \|x_k - y\|^2 - 2\alpha_k (f(x_{k+1}) - f(y)). \end{aligned} \tag{16}$$

Proof (a) We use the formula for the subdifferential of the sum of the three functions f , $(1/2\alpha_k)\|x - x_k\|^2$, and the indicator function of X (cf. Proposition 5.4.6 of [15]), together with the condition that 0 should belong to this subdifferential at the

optimum x_{k+1} . We obtain that Eq. (14) holds if and only if

$$\frac{1}{\alpha_k}(x_k - x_{k+1}) \in \partial f(x_{k+1}) + N_X(x_{k+1}), \tag{17}$$

where $N_X(x_{k+1})$ is the normal cone of X at x_{k+1} [the set of vectors y such that $y'(x - x_{k+1}) \leq 0$ for all $x \in X$, and also the subdifferential of the indicator function of X at x_{k+1} ; see [15], p. 185]. This is true if and only if

$$x_k - x_{k+1} - \alpha_k \tilde{\nabla} f(x_{k+1}) \in N_X(x_{k+1}),$$

for some $\tilde{\nabla} f(x_{k+1}) \in \partial f(x_{k+1})$, which in turn is true if and only if Eq. (15) holds, by the projection theorem.

(b) By writing $\|x_k - y\|^2$ as $\|x_k - x_{k+1} + x_{k+1} - y\|^2$ and expanding, we have

$$\|x_k - y\|^2 = \|x_k - x_{k+1}\|^2 - 2(x_k - x_{k+1})'(y - x_{k+1}) + \|x_{k+1} - y\|^2. \tag{18}$$

Also since from Eq. (17), $\frac{1}{\alpha_k}(x_k - x_{k+1})$ is a subgradient at x_{k+1} of the sum of f and the indicator function of X , we have (using also the assumption $y \in X$)

$$f(x_{k+1}) + \frac{1}{\alpha_k}(x_k - x_{k+1})'(y - x_{k+1}) \leq f(y).$$

Combining this relation with Eq. (18), the result follows. □

Based on part (a) of the preceding proposition, we see that all the iterations (7)–(8), (9)–(10), and (13) can be written in an incremental subgradient format:

(a) Iteration (7)–(8) can be written as

$$z_k = P_X \left(x_k - \alpha_k \tilde{\nabla} f_{i_k}(z_k) \right), \quad x_{k+1} = P_X \left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) \right), \tag{19}$$

(b) Iteration (9)–(10) can be written as

$$z_k = x_k - \alpha_k \tilde{\nabla} f_{i_k}(z_k), \quad x_{k+1} = P_X \left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) \right), \tag{20}$$

(c) Iteration (11)–(12) can be written as

$$z_k = x_k - \alpha_k \tilde{\nabla} h_{i_k}(x_k), \quad x_{k+1} = P_X \left(z_k - \alpha_k \tilde{\nabla} f_{i_k}(x_{k+1}) \right). \tag{21}$$

Using Proposition 1(a), we see that iteration (13) can be written into the form (21), so we will not consider it further. To show this, note that by Proposition 1(b) with

$$f(x) = f_{i_k}(x) + h_{i_k}(x_k) + \tilde{\nabla} h_{i_k}(x_k)'(x - x_k),$$

we may write iteration (13) in the form

$$x_{k+1} = P_X \left(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_{k+1}) - \alpha_k \tilde{\nabla} h_{i_k}(x_k) \right),$$

which is just iteration (21). Note that in all the preceding updates, the subgradient $\tilde{\nabla} h_{i_k}$ can be *any* vector in the subdifferential of h_{i_k} , while the subgradient $\tilde{\nabla} f_{i_k}$ must be a *specific* vector in the subdifferential of f_{i_k} , specified according to Proposition 1(a). Note also that iteration (20) can be written as

$$x_{k+1} = P_X \left(x_k - \alpha_k \tilde{\nabla} F_{i_k}(z_k) \right),$$

and resembles the incremental subgradient method for minimizing over X the cost function

$$F(x) = \sum_{i=1}^m F_i(x) = \sum_{i=1}^m (f_i(x) + h_i(x))$$

[cf. Eq. (5)], the only difference being that the subgradient of F_{i_k} is taken at z_k rather than x_k .

For a convergence analysis, we need to specify the order in which the components $\{f_i, h_i\}$ are chosen for iteration. We consider two possibilities:

- (1) A *cyclic order*, whereby $\{f_i, h_i\}$ are taken up in the fixed deterministic order $1, \dots, m$, so that i_k is equal to $(k \text{ modulo } m) \text{ plus } 1$. A contiguous block of iterations involving $\{f_1, h_1\}, \dots, \{f_m, h_m\}$ in this order and exactly once is called a *cycle*. We assume that the stepsize α_k is constant within a cycle (for all k with $i_k = 1$ we have $\alpha_k = \alpha_{k+1} = \dots = \alpha_{k+m-1}$).
- (2) A *randomized order*, whereby at each iteration a component pair $\{f_i, h_i\}$ is chosen randomly by sampling over all component pairs with a uniform distribution, independently of the past history of the algorithm.⁴

Note that it is essential to include all components in a cycle in the cyclic case, and to sample according to the uniform distribution in the randomized case, for otherwise some components will be sampled more often than others, leading to a bias in the convergence process. For the remainder of the paper, we denote by F^* the optimal value:

$$F^* = \inf_{x \in X} F(x),$$

⁴ Another technique for incremental methods, popular in neural network training practice, is to reshuffle randomly the order of the component functions after each cycle. This alternative order selection scheme leads to convergence, like the preceding two. Moreover, this scheme has the nice property of allocating exactly one computation slot to each component in an m -slot cycle (m incremental iterations). By comparison, choosing components by uniform sampling allocates one computation slot to each component *on the average*, but some components may not get a slot while others may get more than one. A nonzero variance in the number of slots that any fixed component gets within a cycle, may be detrimental to performance, and indicates that reshuffling randomly the order of the component functions after each cycle may work better; this is consistent with experimental observations shared with the author by B. Recht (private communication). However, establishing this fact analytically seems difficult, and remains an open question.

and by X^* the set of optimal solutions (which could be empty):

$$X^* = \{x^* \mid x^* \in X, F(x^*) = F^*\}.$$

Also, for a nonempty closed convex set X , we denote by $\text{dist}(\cdot; X)$ the distance function given by

$$\text{dist}(x; X) = \min_{z \in X} \|x - z\|, \quad x \in \mathfrak{R}^n.$$

In our convergence analysis of Sect. 4, we will use the following well-known theorem (see e.g., [7, 43]). We will use a simpler deterministic version of the theorem in Sect. 3.

Proposition 2 (Supermartingale Convergence Theorem) *Let Y_k, Z_k , and $W_k, k = 0, 1, \dots$, be three sequences of random variables and let $\mathcal{F}_k, k = 0, 1, \dots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k . Suppose that:*

- (1) *The random variables Y_k, Z_k , and W_k are nonnegative, and are functions of the random variables in \mathcal{F}_k .*
- (2) *For each k , we have*

$$E \{Y_{k+1} \mid \mathcal{F}_k\} \leq Y_k - Z_k + W_k.$$

- (3) *There holds, with probability 1, $\sum_{k=0}^\infty W_k < \infty$.*

Then we have $\sum_{k=0}^\infty Z_k < \infty$, and the sequence Y_k converges to a nonnegative random variable Y , with probability 1.

3 Convergence analysis for methods with cyclic order

In this section, we analyze convergence under the cyclic order. We consider a randomized order in the next section. We focus on the sequence $\{x_k\}$ rather than $\{z_k\}$, which need not lie within X in the case of iterations (20) and (21) when $X \neq \mathfrak{R}^n$. In summary, the idea that guides the analysis is to show that the effect of taking subgradients of f_j or h_j at points near x_k (e.g., at z_k rather than at x_k) is inconsequential, and diminishes as the stepsize α_k becomes smaller, as long as some subgradients relevant to the algorithms are uniformly bounded in norm by some constant. In particular, we assume the following throughout this section.

Assumption 1 (For iterations (19) and (20)) *There is a constant $c \in \mathfrak{R}$ such that for all k*

$$\max \left\{ \|\tilde{\nabla} f_{i_k}(z_k)\|, \|\tilde{\nabla} h_{i_k}(z_k)\| \right\} \leq c. \tag{22}$$

Furthermore, for all k that mark the beginning of a cycle (i.e., all $k > 0$ with $i_k = 1$), we have for all $j = 1, \dots, m$,

$$\max \{f_j(x_k) - f_j(z_{k+j-1}), h_j(x_k) - h_j(z_{k+j-1})\} \leq c \|x_k - z_{k+j-1}\|. \tag{23}$$

Assumption 2 (For iteration (21)) There is a constant $c \in \Re$ such that for all k

$$\max \left\{ \|\tilde{\nabla} f_{i_k}(x_{k+1})\|, \|\tilde{\nabla} h_{i_k}(x_k)\| \right\} \leq c. \tag{24}$$

Furthermore, for all k that mark the beginning of a cycle (i.e., all $k > 0$ with $i_k = 1$), we have for all $j = 1, \dots, m$,

$$\max \left\{ f_j(x_k) - f_j(x_{k+j-1}), h_j(x_k) - h_j(x_{k+j-1}) \right\} \leq c \|x_k - x_{k+j-1}\|, \tag{25}$$

$$f_j(x_{k+j-1}) - f_j(x_{k+j}) \leq c \|x_{k+j-1} - x_{k+j}\|. \tag{26}$$

Note that conditions (23) and (25) are satisfied if for each j and k , there is a subgradient of f_j at x_k and a subgradient of h_j at x_k , whose norms are bounded by c . Conditions that imply the preceding assumptions are that:

- (a) For algorithm (19): f_i and h_i are Lipschitz continuous over X .
- (b) For algorithms (20) and (21): f_i and h_i are Lipschitz continuous over \Re^n .
- (c) For all algorithms (19), (20), and (21): f_i and h_i are polyhedral [this is a special case of (a) and (b)].
- (d) The sequences $\{x_k\}$ and $\{z_k\}$ are bounded [since then f_i and h_i , being real-valued and convex, are Lipschitz continuous over any bounded set that contains $\{x_k\}$ and $\{z_k\}$ (see e.g., [15], Proposition 5.4.2)].

The following proposition provides a key estimate.

Proposition 3 Let $\{x_k\}$ be the sequence generated by any one of the algorithms (19)–(21), with a cyclic order of component selection. Then for all $y \in X$ and all k that mark the beginning of a cycle (i.e., all k with $i_k = 1$), we have

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k (F(x_k) - F(y)) + \alpha_k^2 \beta m^2 c^2, \tag{27}$$

where $\beta = \frac{1}{m} + 4$ in the case of (19) and (20), and $\beta = \frac{5}{m} + 4$ in the case of (21).

Proof We first prove the result for algorithms (19) and (20), and then indicate the modifications necessary for algorithm (21). Using Proposition 1(b), we have for all $y \in X$ and k ,

$$\|z_k - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k (f_{i_k}(z_k) - f_{i_k}(y)). \tag{28}$$

Also, using the nonexpansion property of the projection [i.e., $\|P_X(u) - P_X(v)\| \leq \|u - v\|$ for all $u, v \in \Re^n$], the definition of subgradient, and Eq. (22), we obtain for all $y \in X$ and k ,

$$\begin{aligned} \|x_{k+1} - y\|^2 &= \|P_X \left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) \right) - y\|^2 \\ &\leq \|z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k) - y\|^2 \\ &\leq \|z_k - y\|^2 - 2\alpha_k \tilde{\nabla} h_{i_k}(z_k)'(z_k - y) + \alpha_k^2 \left\| \tilde{\nabla} h_{i_k}(z_k) \right\|^2 \\ &\leq \|z_k - y\|^2 - 2\alpha_k (h_{i_k}(z_k) - h_{i_k}(y)) + \alpha_k^2 c^2. \end{aligned} \tag{29}$$

Combining Eqs. (28) and (29), and using the definition $F_j = f_j + h_j$, we have

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (f_{i_k}(z_k) + h_{i_k}(z_k) - f_{i_k}(y) - h_{i_k}(y)) + \alpha_k^2 c^2 \\ &= \|x_k - y\|^2 - 2\alpha_k (F_{i_k}(z_k) - F_{i_k}(y)) + \alpha_k^2 c^2. \end{aligned} \tag{30}$$

Let now k mark the beginning of a cycle (i.e., $i_k = 1$). Then at iteration $k + j - 1$, $j = 1, \dots, m$, the selected components are $\{f_j, h_j\}$, in view of the assumed cyclic order. We may thus replicate the preceding inequality with k replaced by $k + 1, \dots, k + m - 1$, and add to obtain

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \sum_{j=1}^m (F_j(z_{k+j-1}) - F_j(y)) + m\alpha_k^2 c^2,$$

or equivalently, since $F = \sum_{j=1}^m F_j$,

$$\begin{aligned} \|x_{k+m} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (F(x_k) - F(y)) + m\alpha_k^2 c^2 \\ &\quad + 2\alpha_k \sum_{j=1}^m (F_j(x_k) - F_j(z_{k+j-1})). \end{aligned} \tag{31}$$

The remainder of the proof deals with appropriately bounding the last term above.

From Eq. (23), we have for $j = 1, \dots, m$,

$$F_j(x_k) - F_j(z_{k+j-1}) \leq 2c \|x_k - z_{k+j-1}\|. \tag{32}$$

We also have

$$\|x_k - z_{k+j-1}\| \leq \|x_k - x_{k+1}\| + \dots + \|x_{k+j-2} - x_{k+j-1}\| + \|x_{k+j-1} - z_{k+j-1}\|, \tag{33}$$

and by the definition of the algorithms (19) and (20), the nonexpansion property of the projection, and Eq. (22), each of the terms in the right-hand side above is bounded by $2\alpha_k c$, except for the last, which is bounded by $\alpha_k c$. Thus Eq. (33) yields $\|x_k - z_{k+j-1}\| \leq \alpha_k (2j - 1)c$, which together with Eq. (32), shows that

$$F_j(x_k) - F_j(z_{k+j-1}) \leq 2\alpha_k c^2 (2j - 1). \tag{34}$$

Combining Eqs. (31) and (34), we have

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k (F(x_k) - F(y)) + m\alpha_k^2 c^2 + 4\alpha_k^2 c^2 \sum_{j=1}^m (2j - 1),$$

and finally

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k (F(x_k) - F(y)) + m\alpha_k^2 c^2 + 4\alpha_k^2 c^2 m^2,$$

which is of the form (27) with $\beta = \frac{1}{m} + 4$.

For the algorithm (21), a similar argument goes through using Assumption 2. In place of Eq. (28), using the nonexpansion property of the projection, the definition of subgradient, and Eq. (24), we obtain for all $y \in X$ and $k \geq 0$,

$$\|z_k - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k (h_{i_k}(x_k) - h_{i_k}(y)) + \alpha_k^2 c^2, \tag{35}$$

while in place of Eq. (29), using Proposition 1(b), we have

$$\|x_{k+1} - y\|^2 \leq \|z_k - y\|^2 - 2\alpha_k (f_{i_k}(x_{k+1}) - f_{i_k}(y)). \tag{36}$$

Combining these equations, in analogy with Eq. (30), we obtain

$$\begin{aligned} \|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (f_{i_k}(x_{k+1}) + h_{i_k}(x_k) - f_{i_k}(y) - h_{i_k}(y)) + \alpha_k^2 c^2 \\ &= \|x_k - y\|^2 - 2\alpha_k (F_{i_k}(x_k) - F_{i_k}(y)) + \alpha_k^2 c^2 + 2\alpha_k (f_{i_k}(x_k) - f_{i_k}(x_{k+1})). \end{aligned} \tag{37}$$

As earlier, we let k mark the beginning of a cycle (i.e., $i_k = 1$). We replicate the preceding inequality with k replaced by $k + 1, \dots, k + m - 1$, and add to obtain [in analogy with Eq. (31)]

$$\begin{aligned} \|x_{k+m} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k (F(x_k) - F(y)) + m\alpha_k^2 c^2 \\ &\quad + 2\alpha_k \sum_{j=1}^m (F_j(x_k) - F_j(x_{k+j-1})) + 2\alpha_k \sum_{j=1}^m (f_j(x_{k+j-1}) - f_j(x_{k+j})). \end{aligned} \tag{38}$$

[Note that relative to Eq. (31), the preceding equation contains an extra last term, which results from a corresponding extra term in Eq. (37) relative to Eq. (30). This accounts for the difference in the value of β in the statement of the proposition.]

We now bound the two sums in Eq. (38), using Assumption 2. From Eq. (25), we have

$$\begin{aligned} F_j(x_k) - F_j(x_{k+j-1}) &\leq 2c \|x_k - x_{k+j-1}\| \leq 2c (\|x_k - x_{k+1}\| + \dots + \|x_{k+j-2} - x_{k+j-1}\|), \end{aligned}$$

and since by Eq. (24) and the definition of the algorithm, each of the norm terms in the right-hand side above is bounded by $2\alpha_k c$,

$$F_j(x_k) - F_j(x_{k+j-1}) \leq 4\alpha_k c^2 (j - 1).$$

Also from Eqs. (24) and (47), and the nonexpansion property of the projection, we have

$$f_j(x_{k+j-1}) - f_j(x_{k+j}) \leq c \|x_{k+j-1} - x_{k+j}\| \leq 2\alpha_k c^2.$$

Combining the preceding relations and adding, we obtain

$$\begin{aligned} & 2\alpha_k \sum_{j=1}^m (F_j(x_k) - F_j(x_{k+j-1})) + 2\alpha_k \sum_{j=1}^m (f_j(x_{k+j-1}) - f_j(x_{k+j})) \\ & \leq 8\alpha_k^2 c^2 \sum_{j=1}^m (j - 1) + 4\alpha_k^2 c^2 m \\ & = 4\alpha_k^2 c^2 m^2 + 4\alpha_k^2 c^2 m \\ & = \left(4 + \frac{4}{m}\right) \alpha_k^2 c^2 m^2, \end{aligned}$$

which together with Eq. (38), yields Eq. (27) with $\beta = 4 + \frac{5}{m}$. □

Among other things, Proposition 3 guarantees that with a cyclic order, given the iterate x_k at the start of a cycle and any point $y \in X$ having lower cost than x_k , the algorithm yields a point x_{k+m} at the end of the cycle that will be closer to y than x_k , provided the stepsize α_k is sufficiently small [less than $2(F(x_k) - F(y)) / \beta m^2 c^2$]. In particular, for any $\epsilon > 0$ and assuming that there exists an optimal solution x^* , either we are within $\frac{\alpha_k \beta m^2 c^2}{2} + \epsilon$ of the optimal value,

$$F(x_k) \leq F(x^*) + \frac{\alpha_k \beta m^2 c^2}{2} + \epsilon,$$

or else the squared distance to x^* will be strictly decreased by at least $2\alpha_k \epsilon$,

$$\|x_{k+m} - x^*\|^2 < \|x_k - x^*\|^2 - 2\alpha_k \epsilon.$$

Thus, using this argument, we can provide convergence results for various stepsize rules, and this is done in the next two subsections.

3.1 Convergence within an error bound for a constant stepsize

For a constant stepsize ($\alpha_k \equiv \alpha$), convergence can be established to a neighborhood of the optimum, which shrinks to 0 as $\alpha \rightarrow 0$. We show this in the following proposition.

Proposition 4 *Let $\{x_k\}$ be the sequence generated by any one of the algorithms (19)–(21), with a cyclic order of component selection, and let the stepsize α_k be fixed at some positive constant α .*

(a) If $F^* = -\infty$, then

$$\liminf_{k \rightarrow \infty} F(x_k) = F^*.$$

(b) If $F^* > -\infty$, then

$$\liminf_{k \rightarrow \infty} F(x_k) \leq F^* + \frac{\alpha\beta m^2 c^2}{2},$$

where c and β are the constants of Proposition 3.

Proof We prove (a) and (b) simultaneously. If the result does not hold, there must exist an $\epsilon > 0$ such that

$$\liminf_{k \rightarrow \infty} F(x_{km}) - \frac{\alpha\beta m^2 c^2}{2} - 2\epsilon > F^*.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k \rightarrow \infty} F(x_{km}) - \frac{\alpha\beta m^2 c^2}{2} - 2\epsilon \geq F(\hat{y}),$$

and let k_0 be large enough so that for all $k \geq k_0$, we have

$$F(x_{km}) \geq \liminf_{k \rightarrow \infty} F(x_{km}) - \epsilon.$$

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

$$F(x_{km}) - F(\hat{y}) \geq \frac{\alpha\beta m^2 c^2}{2} + \epsilon.$$

Using Proposition 3 for the case where $y = \hat{y}$ together with the above relation, we obtain for all $k \geq k_0$,

$$\begin{aligned} \|x_{(k+1)m} - \hat{y}\|^2 &\leq \|x_{km} - \hat{y}\|^2 - 2\alpha (F(x_{km}) - F(\hat{y})) + \beta\alpha^2 m^2 c^2 \\ &\leq \|x_{km} - \hat{y}\|^2 - 2\alpha\epsilon. \end{aligned}$$

This relation implies that for all $k \geq k_0$,

$$\|x_{(k+1)m} - \hat{y}\|^2 \leq \|x_{(k-1)m} - \hat{y}\|^2 - 4\alpha\epsilon \leq \dots \leq \|x_{k_0} - \hat{y}\|^2 - 2(k + 1 - k_0)\alpha\epsilon,$$

which cannot hold for k sufficiently large—a contradiction. □

The next proposition gives an estimate of the number of iterations needed to guarantee a given level of optimality up to the threshold tolerance $\alpha\beta m^2 c^2/2$ given in the preceding proposition.

Proposition 5 *Let $\{x_k\}$ be a sequence generated as in Proposition 4. Then for $\epsilon > 0$, we have*

$$\min_{0 \leq k \leq N} F(x_k) \leq F^* + \frac{\alpha\beta m^2 c^2 + \epsilon}{2}, \tag{39}$$

where N is given by

$$N = m \left\lceil \frac{\text{dist}(x_0; X^*)^2}{\alpha\epsilon} \right\rceil. \tag{40}$$

Proof Assume, to arrive at a contradiction, that Eq. (39) does not hold, so that for all k with $0 \leq km \leq N$, we have

$$F(x_{km}) > F^* + \frac{\alpha\beta m^2 c^2 + \epsilon}{2}.$$

By using this relation in Proposition 3 with α_k replaced by α and y equal to the vector of X^* that is at minimum distance from x_{km} , we obtain for all k with $0 \leq km \leq N$,

$$\begin{aligned} \text{dist}(x_{(k+1)m}; X^*)^2 &\leq \text{dist}(x_{km}; X^*)^2 - 2\alpha(F(x_{km}) - F^*) + \alpha^2\beta m^2 c^2 \\ &\leq \text{dist}(x_{km}; X^*)^2 - (\alpha^2\beta m^2 c^2 + \alpha\epsilon) + \alpha^2\beta m^2 c^2 \\ &= \text{dist}(x_{km}; X^*)^2 - \alpha\epsilon. \end{aligned}$$

Adding the above inequalities for $k = 0, \dots, \frac{N}{m}$, we obtain

$$\text{dist}(x_{N+m}; X^*)^2 \leq \text{dist}(x_0; X^*)^2 - \left(\frac{N}{m} + 1\right)\alpha\epsilon,$$

so that

$$\left(\frac{N}{m} + 1\right)\alpha\epsilon \leq \text{dist}(x_0; X^*)^2,$$

which contradicts the definition of N . □

According to Proposition 5, to achieve a cost function value within $O(\epsilon)$ of the optimal, the term $\alpha\beta m^2 c^2$ must also be of order $O(\epsilon)$, so α must be of order $O(\epsilon/m^2 c^2)$, and from Eq. (40), the number of necessary iterations N is $O(m^3 c^2/\epsilon^2)$, and the number of necessary cycles is $O((mc)^2/\epsilon^2)$. This is the same type of estimate as for the nonincremental subgradient method [i.e., $O(1/\epsilon^2)$, counting a cycle as one iteration of the nonincremental method, and viewing mc as a Lipschitz constant for the entire cost function F], and does not reveal any advantage for the incremental methods given here. However, in the next section, we demonstrate a much more favorable iteration complexity estimate for the incremental methods that use a randomized order of component selection.

3.2 Exact convergence for a diminishing stepsize

We also obtain an exact convergence result for the case where the stepsize α_k diminishes to zero, but satisfies $\sum_{k=0}^{\infty} \alpha_k = \infty$ (so that the method can “travel” infinitely far if necessary).

Proposition 6 *Let $\{x_k\}$ be the sequence generated by any one of the algorithms (19)–(21), with a cyclic order of component selection, and let the stepsize α_k satisfy*

$$\lim_{k \rightarrow \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then,

$$\liminf_{k \rightarrow \infty} F(x_k) = F^*.$$

Furthermore, if X^* is nonempty and

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

then $\{x_k\}$ converges to some $x^* \in X^*$.

Proof For the first part, it will be sufficient to show that $\liminf_{k \rightarrow \infty} F(x_{km}) = F^*$. Assume, to arrive at a contradiction, that there exists an $\epsilon > 0$ such that

$$\liminf_{k \rightarrow \infty} F(x_{km}) - 2\epsilon > F^*.$$

Then there exists a point $\hat{y} \in X$ such that

$$\liminf_{k \rightarrow \infty} F(x_{km}) - 2\epsilon > F(\hat{y}).$$

Let k_0 be large enough so that for all $k \geq k_0$, we have

$$F(x_{km}) \geq \liminf_{k \rightarrow \infty} F(x_{km}) - \epsilon.$$

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

$$F(x_{km}) - F(\hat{y}) > \epsilon.$$

By setting $y = \hat{y}$ in Proposition 3, and by using the above relation, we have for all $k \geq k_0$,

$$\begin{aligned} \|x_{(k+1)m} - \hat{y}\|^2 &\leq \|x_{km} - \hat{y}\|^2 - 2\alpha_{km}\epsilon + \beta\alpha_{km}^2 m^2 c^2 \\ &= \|x_{km} - \hat{y}\|^2 - \alpha_{km} \left(2\epsilon - \beta\alpha_{km} m^2 c^2\right). \end{aligned}$$

Since $\alpha_k \rightarrow 0$, without loss of generality, we may assume that k_0 is large enough so that

$$2\epsilon - \beta\alpha_k m^2 c^2 \geq \epsilon, \quad \forall k \geq k_0.$$

Therefore for all $k \geq k_0$, we have

$$\|x_{(k+1)m} - \hat{y}\|^2 \leq \|x_{km} - \hat{y}\|^2 - \alpha_{km}\epsilon \leq \dots \leq \|x_{k_0m} - \hat{y}\|^2 - \epsilon \sum_{\ell=k_0}^k \alpha_{\ell m},$$

which cannot hold for k sufficiently large. Hence $\liminf_{k \rightarrow \infty} F(x_{km}) = F^*$.

To prove the second part of the proposition, note that from Proposition 3, for every $x^* \in X^*$ and $k \geq 0$ we have

$$\|x_{(k+1)m} - x^*\|^2 \leq \|x_{km} - x^*\|^2 - 2\alpha_{km} (F(x_{km}) - F(x^*)) + \alpha_{km}^2 \beta m^2 c^2. \quad (41)$$

From the Supermartingale Convergence Theorem (Proposition 2) and the hypothesis $\sum_{k=0}^\infty \alpha_k^2 < \infty$, we see that $\{\|x_{km} - x^*\|\}$ converges for every $x^* \in X^*$.⁵ Since then $\{x_{km}\}$ is bounded, it has a limit point $\bar{x} \in X$ that satisfies

$$F(\bar{x}) = \liminf_{k \rightarrow \infty} F(x_{km}) = F^*.$$

This implies that $\bar{x} \in X^*$, so it follows that $\{\|x_{km} - \bar{x}\|\}$ converges, and that the entire sequence $\{x_{km}\}$ converges to \bar{x} (since \bar{x} is a limit point of $\{x_{km}\}$).

Finally, to show that the entire sequence $\{x_k\}$ also converges to \bar{x} , note that from Eqs. (22) and (24), and the form of the iterations (19)–(21), we have $\|x_{k+1} - x_k\| \leq 2\alpha_k c \rightarrow 0$. Since $\{x_{km}\}$ converges to \bar{x} , it follows that $\{x_k\}$ also converges to \bar{x} . \square

4 Convergence analysis for methods with randomized order

In this section, we analyze our algorithms for the randomized component selection order and a constant stepsize α . The randomized versions of iterations (19), (20), and (21), are

⁵ Actually we use here a deterministic version/special case of the theorem, where Y_k, Z_k , and W_k are nonnegative scalar sequences satisfying $Y_{k+1} \leq Y_k - Z_k + W_k$ with $\sum_{k=0}^\infty W_k < \infty$. Then the sequence Y_k must converge. This version is given with proof in many sources, including [7] (Lemma 3.4), and [8] (Lemma 1).

$$z_k = P_X \left(x_k - \alpha \tilde{\nabla} f_{\omega_k}(z_k) \right), \quad x_{k+1} = P_X \left(z_k - \alpha \tilde{\nabla} h_{\omega_k}(z_k) \right), \quad (42)$$

$$z_k = x_k - \alpha \tilde{\nabla} f_{\omega_k}(z_k), \quad x_{k+1} = P_X \left(z_k - \alpha \tilde{\nabla} h_{\omega_k}(z_k) \right), \quad (43)$$

$$z_k = x_k - \alpha_k \tilde{\nabla} h_{\omega_k}(x_k), \quad x_{k+1} = P_X \left(z_k - \alpha_k \tilde{\nabla} f_{\omega_k}(x_{k+1}) \right), \quad (44)$$

respectively, where $\{\omega_k\}$ is a sequence of random variables, taking values from the index set $\{1, \dots, m\}$.

We assume the following throughout the present section.

Assumption 3 (For iterations (42) and (43)) (a) $\{\omega_k\}$ is a sequence of random variables, each uniformly distributed over $\{1, \dots, m\}$, and such that for each k , ω_k is independent of the past history $\{x_k, z_{k-1}, x_{k-1}, \dots, z_0, x_0\}$.

(b) There is a constant $c \in \Re$ such that for all k , we have with probability 1

$$\max \left\{ \|\tilde{\nabla} f_i(z_k^i)\|, \|\tilde{\nabla} h_i(z_k^i)\| \right\} \leq c, \quad \forall i = 1, \dots, m, \quad (45)$$

$$\max \left\{ f_i(x_k) - f_i(z_k^i), h_i(x_k) - h_i(z_k^i) \right\} \leq c \|x_k - z_k^i\|, \quad \forall i = 1, \dots, m, \quad (46)$$

where z_k^i is the result of the proximal iteration, starting at x_k if ω_k would be i , i.e.,

$$z_k^i = \arg \min_{x \in X} \left\{ f_i(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\},$$

in the case of iteration (42), and

$$z_k^i = \arg \min_{x \in \Re^n} \left\{ f_i(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\},$$

in the case of iteration (43).

Assumption 4 (For iteration (44)) (a) $\{\omega_k\}$ is a sequence of random variables, each uniformly distributed over $\{1, \dots, m\}$, and such that for each k , ω_k is independent of the past history $\{x_k, z_{k-1}, x_{k-1}, \dots, z_0, x_0\}$.

(b) There is a constant $c \in \Re$ such that for all k , we have with probability 1

$$\max \left\{ \|\tilde{\nabla} f_i(x_{k+1}^i)\|, \|\tilde{\nabla} h_i(x_k)\| \right\} \leq c, \quad \forall i = 1, \dots, m, \quad (47)$$

$$f_i(x_k) - f_i(x_{k+1}^i) \leq c \|x_k - x_{k+1}^i\|, \quad \forall i = 1, \dots, m, \quad (48)$$

where x_{k+1}^i is the result of the iteration, starting at x_k if ω_k would be i , i.e.,

$$x_{k+1}^i = P_X \left(z_k^i - \alpha_k \tilde{\nabla} f_i(x_{k+1}^i) \right),$$

with

$$z_k^i = x_k - \alpha_k \tilde{\nabla} h_i(x_k).$$

Note that condition (46) is satisfied if there exist subgradients of f_i and h_i at x_k with norms less than or equal to c . Thus the conditions (45) and (46) are similar, the main difference being that the first applies to “slopes” of f_i and h_i at z_k^i while the second applies to the “slopes” of f_i and h_i at x_k . There is an analogous similarity between conditions (47) and (48). As in the case of Assumptions 1 and 2, these conditions are guaranteed by Lipschitz continuity assumptions on f_i and h_i . We will first deal with the case of a constant stepsize, and then consider the case of a diminishing stepsize.

Proposition 7 *Let $\{x_k\}$ be the sequence generated by one of the randomized incremental methods (42)–(44), and let the stepsize α_k be fixed at some positive constant α .*

(a) *If $F^* = -\infty$, then with probability 1*

$$\inf_{k \geq 0} F(x_k) = F^*.$$

(b) *If $F^* > -\infty$, then with probability 1*

$$\inf_{k \geq 0} F(x_k) \leq F^* + \frac{\alpha \beta m c^2}{2},$$

where $\beta = 5$.

Proof Consider first algorithms (42) and (43). By adapting the proof argument of Proposition 3 with F_{i_k} replaced by F_{ω_k} [cf. Eq. (30)], we have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha (F_{\omega_k}(z_k) - F_{\omega_k}(y)) + \alpha^2 c^2, \quad \forall y \in X, \quad k \geq 0.$$

By taking the conditional expectation with respect to $\mathcal{F}_k = \{x_k, z_{k-1}, \dots, z_0, x_0\}$, and using the fact that ω_k takes the values $i = 1, \dots, m$ with equal probability $1/m$, we obtain for all $y \in X$ and k ,

$$\begin{aligned} E \left\{ \|x_{k+1} - y\|^2 \mid \mathcal{F}_k \right\} &\leq \|x_k - y\|^2 - 2\alpha E \left\{ F_{\omega_k}(z_k) - F_{\omega_k}(y) \mid \mathcal{F}_k \right\} + \alpha^2 c^2 \\ &= \|x_k - y\|^2 - \frac{2\alpha}{m} \sum_{i=1}^m \left(F_i(z_k^i) - F_i(y) \right) + \alpha^2 c^2 \\ &= \|x_k - y\|^2 - \frac{2\alpha}{m} (F(x_k) - F(y)) + \frac{2\alpha}{m} \sum_{i=1}^m \left(F_i(x_k) - F_i(z_k^i) \right) + \alpha^2 c^2. \end{aligned} \tag{49}$$

By using Eqs. (45) and (46),

$$\sum_{i=1}^m \left(F_i(x_k) - F_i(z_k^i) \right) \leq 2c \sum_{i=1}^m \|x_k - z_k^i\| = 2c\alpha \sum_{i=1}^m \|\tilde{\nabla} f_i(z_k^i)\| \leq 2m\alpha c^2.$$

By combining the preceding two relations, we obtain

$$\begin{aligned}
 E\{\|x_{k+1} - y\|^2 \mid \mathcal{F}_k\} &\leq \|x_k - y\|^2 - \frac{2\alpha}{m} (F(x_k) - F(y)) + 4\alpha^2 c^2 + \alpha^2 c^2 \\
 &= \|x_k - y\|^2 - \frac{2\alpha}{m} (F(x_k) - F(y)) + \beta\alpha^2 c^2, \tag{50}
 \end{aligned}$$

where $\beta = 5$.

The preceding equation holds also for algorithm (44). To see this note that Eq. (37) yields for all $y \in X$

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha (F_{\omega_k}(x_k) - F_{\omega_k}(y)) + \alpha^2 c^2 + 2\alpha (f_{\omega_k}(x_k) - f_{\omega_k}(x_{k+1})). \tag{51}$$

and similar to Eq. (49), we obtain

$$\begin{aligned}
 E\{\|x_{k+1} - y\|^2 \mid \mathcal{F}_k\} &\leq \|x_k - y\|^2 - \frac{2\alpha}{m} (F(x_k) - F(y)) \\
 &\quad + \frac{2\alpha}{m} \sum_{i=1}^m (f_i(x_k) - f_i(x_{k+1}^i)) + \alpha^2 c^2. \tag{52}
 \end{aligned}$$

From Eq. (48), we have

$$f_i(x_k) - f_i(x_{k+1}^i) \leq c\|x_k - x_{k+1}^i\|,$$

and from Eq. (47) and the nonexpansion property of the projection,

$$\|x_k - x_{k+1}^i\| \leq \|x_k - z_k^i + \alpha \tilde{\nabla} f_i(x_{k+1}^i)\| = \|x_k - x_k + \alpha \tilde{\nabla} h_i(x_k) + \alpha \tilde{\nabla} f_i(x_{k+1}^i)\| \leq 2\alpha c.$$

Combining the preceding inequalities, we obtain Eq. (50) with $\beta = 5$.

Let us fix a positive scalar γ , consider the level set L_γ defined by

$$L_\gamma = \begin{cases} \left\{ x \in X \mid F(x) < -\gamma + 1 + \frac{\alpha\beta mc^2}{2} \right\} & \text{if } F^* = -\infty, \\ \left\{ x \in X \mid F(x) < F^* + \frac{2}{\gamma} + \frac{\alpha\beta mc^2}{2} \right\} & \text{if } F^* > -\infty, \end{cases}$$

and let $y_\gamma \in X$ be such that

$$F(y_\gamma) = \begin{cases} -\gamma & \text{if } F^* = -\infty, \\ F^* + \frac{1}{\gamma} & \text{if } F^* > -\infty, \end{cases}$$

Note that $y_\gamma \in L_\gamma$ by construction. Define a new process $\{\hat{x}_k\}$ that is identical to $\{x_k\}$, except that once x_k enters the level set L_γ , the process terminates with $\hat{x}_k = y_\gamma$. We will now argue that for any fixed γ , $\{\hat{x}_k\}$ (and hence also $\{x_k\}$) will eventually enter L_γ , which will prove both parts (a) and (b).

Using Eq. (50) with $y = y_\gamma$, we have

$$E \left\{ \|\hat{x}_{k+1} - y_\gamma\|^2 \mid \mathcal{F}_k \right\} \leq \|\hat{x}_k - y_\gamma\|^2 - \frac{2\alpha}{m} (F(\hat{x}_k) - F(y_\gamma)) + \beta\alpha^2c^2,$$

from which

$$E \left\{ \|\hat{x}_{k+1} - y_\gamma\|^2 \mid \mathcal{F}_k \right\} \leq \|\hat{x}_k - y_\gamma\|^2 - v_k, \tag{53}$$

where

$$v_k = \begin{cases} \frac{2\alpha}{m} (F(\hat{x}_k) - F(y_\gamma)) - \beta\alpha^2c^2 & \text{if } \hat{x}_k \notin L_\gamma, \\ 0 & \text{if } \hat{x}_k = y_\gamma, \end{cases}$$

The idea of the subsequent argument is to show that as long as $\hat{x}_k \notin L_\gamma$, the scalar v_k (which is a measure of progress) is strictly positive and bounded away from 0.

(a) Let $F^* = -\infty$. Then if $\hat{x}_k \notin L_\gamma$, we have

$$\begin{aligned} v_k &= \frac{2\alpha}{m} (F(\hat{x}_k) - F(y_\gamma)) - \beta\alpha^2c^2 \\ &\geq \frac{2\alpha}{m} \left(-\gamma + 1 + \frac{\alpha\beta mc^2}{2} + \gamma \right) - \beta\alpha^2c^2 \\ &= \frac{2\alpha}{m}. \end{aligned}$$

Since $v_k = 0$ for $\hat{x}_k \in L_\gamma$, we have $v_k \geq 0$ for all k , and by Eq. (53) and the Supermartingale Convergence Theorem (cf. Proposition 2), we obtain $\sum_{k=0}^\infty v_k < \infty$ implying that $\hat{x}_k \in L_\gamma$ for sufficiently large k , with probability 1. Therefore, in the original process we have with probability 1

$$\inf_{k \geq 0} F(x_k) \leq -\gamma + 1 + \frac{\alpha\beta mc^2}{2}.$$

Letting $\gamma \rightarrow \infty$, we obtain $\inf_{k \geq 0} F(x_k) = -\infty$ with probability 1.

(b) Let $F^* > -\infty$. Then if $\hat{x}_k \notin L_\gamma$, we have

$$\begin{aligned} v_k &= \frac{2\alpha}{m} (F(\hat{x}_k) - F(y_\gamma)) - \beta\alpha^2c^2 \\ &\geq \frac{2\alpha}{m} \left(F^* + \frac{2}{\gamma} + \frac{\alpha\beta mc^2}{2} - F^* - \frac{1}{\gamma} \right) - \beta\alpha^2c^2 \\ &= \frac{2\alpha}{m\gamma}. \end{aligned}$$

Hence, $v_k \geq 0$ for all k , and by the Supermartingale Convergence Theorem, we have $\sum_{k=0}^\infty v_k < \infty$ implying that $\hat{x}_k \in L_\gamma$ for sufficiently large k , so that in the original

process,

$$\inf_{k \geq 0} F(x_k) \leq F^* + \frac{2}{\gamma} + \frac{\alpha\beta mc^2}{2}$$

with probability 1. Letting $\gamma \rightarrow \infty$, we obtain $\inf_{k \geq 0} F(x_k) \leq F^* + \alpha\beta mc^2/2$. \square

4.1 Error bound for a constant stepsize

By comparing Proposition 7(b) with Proposition 4(b), we see that when $F^* > -\infty$ and the stepsize α is constant, the randomized methods (42), (43), and (44), have a better error bound (by a factor m) than their nonrandomized counterparts. It is important to note that the bound of Proposition 4(b) is tight in the sense that for a bad problem/cyclic order we have $\liminf_{k \rightarrow \infty} F(x_k) - F^* = O(\alpha m^2 c^2)$ (an example where $f_i \equiv 0$ is given in p. 514 of [5]). By contrast the randomized method will get to within $O(\alpha mc^2)$ with probability 1 for any problem, according to Proposition 7(b). Thus the randomized order provides a worst-case performance advantage over the cyclic order: we do not run the risk of choosing by accident a bad cyclic order. Note, however, that this assessment is relevant to asymptotic convergence; the cyclic and randomized order algorithms appear to perform comparably when far from convergence for the same stepsize α .

A related convergence rate result is provided by the following proposition, which should be compared with Proposition 5 for the nonrandomized methods.

Proposition 8 *Assume that X^* is nonempty. Let $\{x_k\}$ be a sequence generated as in Proposition 7. Then for any positive scalar ϵ , we have with probability 1*

$$\min_{0 \leq k \leq N} F(x_k) \leq F^* + \frac{\alpha\beta mc^2 + \epsilon}{2}, \quad (54)$$

where N is a random variable with

$$E\{N\} \leq m \frac{\text{dist}(x_0; X^*)^2}{\alpha\epsilon}. \quad (55)$$

Proof Let \hat{y} be some fixed vector in X^* . Define a new process $\{\hat{x}_k\}$ which is identical to $\{x_k\}$ except that once x_k enters the level set

$$L = \left\{ x \in X \mid F(x) < F^* + \frac{\alpha\beta mc^2 + \epsilon}{2} \right\},$$

the process $\{\hat{x}_k\}$ terminates at \hat{y} . Similar to the proof of Proposition 7 [cf. Eq. (50) with y being the closest point of \hat{x}_k in X^*], for the process $\{\hat{x}_k\}$ we obtain for all k ,

$$\begin{aligned} E \left\{ \text{dist}(\hat{x}_{k+1}; X^*)^2 \mid \mathcal{F}_k \right\} &\leq E \left\{ \|\hat{x}_{k+1} - y\|^2 \mid \mathcal{F}_k \right\} \\ &\leq \text{dist}(\hat{x}_k; X^*)^2 - \frac{2\alpha}{m} (F(\hat{x}_k) - F^*) + \beta\alpha^2c^2 \\ &= \text{dist}(\hat{x}_k; X^*)^2 - v_k, \end{aligned} \tag{56}$$

where $\mathcal{F}_k = \{x_k, z_{k-1}, \dots, z_0, x_0\}$ and

$$v_k = \begin{cases} \frac{2\alpha}{m} (F(\hat{x}_k) - F^*) - \beta\alpha^2c^2 & \text{if } \hat{x}_k \notin L, \\ 0 & \text{otherwise.} \end{cases}$$

In the case where $\hat{x}_k \notin L$, we have

$$v_k \geq \frac{2\alpha}{m} \left(F^* + \frac{\alpha\beta mc^2 + \epsilon}{2} - F^* \right) - \beta\alpha^2c^2 = \frac{\alpha\epsilon}{m}. \tag{57}$$

By the Supermartingale Convergence Theorem (cf. Proposition 2), from Eq. (56) we have

$$\sum_{k=0}^{\infty} v_k < \infty$$

with probability 1, so that $v_k = 0$ for all $k \geq N$, where N is a random variable. Hence $\hat{x}_N \in L$ with probability 1, implying that in the original process we have

$$\min_{0 \leq k \leq N} F(x_k) \leq F^* + \frac{\alpha\beta mc^2 + \epsilon}{2}$$

with probability 1. Furthermore, by taking the total expectation in Eq. (56), we obtain for all k ,

$$E \left\{ \text{dist}(\hat{x}_{k+1}; X^*)^2 \right\} \leq E \left\{ \text{dist}(\hat{x}_k; X^*)^2 \right\} - E\{v_k\} \leq \text{dist}(\hat{x}_0; X^*)^2 - E \left\{ \sum_{j=0}^k v_j \right\},$$

where in the last inequality we use the facts $\hat{x}_0 = x_0$ and $E \left\{ \text{dist}(\hat{x}_0; X^*)^2 \right\} = \text{dist}(\hat{x}_0; X^*)^2$. Therefore, letting $k \rightarrow \infty$, and using the definition of v_k and Eq. (57),

$$\text{dist}(\hat{x}_0; X^*)^2 \geq E \left\{ \sum_{k=0}^{\infty} v_k \right\} = E \left\{ \sum_{k=0}^{N-1} v_k \right\} \geq E \left\{ \frac{N\alpha\epsilon}{m} \right\} = \frac{\alpha\epsilon}{m} E\{N\}.$$

□

A comparison of Propositions 5 and 8 again suggests an advantage for the randomized order: compared to the cyclic order, it achieves a much smaller error tolerance (a factor of m), in the same *expected* number of iterations. Note, however, that the preceding assessment is based on upper bound estimates, which may not be sharp on a given problem [although the bound of Proposition 4(b) is tight with a worst-case problem selection as mentioned earlier; see [5], p. 514]. Moreover, the comparison based on worst-case values versus expected values may not be strictly valid. In particular, while Proposition 5 provides an upper bound estimate on N , Proposition 8 provides an upper bound estimate on $E\{N\}$, which is not quite the same.

4.2 Exact convergence for a diminishing stepsize rule

We finally consider the case of a diminishing stepsize rule and obtain an exact convergence result similar to Proposition 6 for the case of a randomized order selection.

Proposition 9 *Let $\{x_k\}$ be the sequence generated by one of the randomized incremental methods (42)–(44), and let the stepsize α_k satisfy*

$$\lim_{k \rightarrow \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, with probability 1,

$$\liminf_{k \rightarrow \infty} F(x_k) = F^*.$$

Furthermore, if X^ is nonempty and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, then $\{x_k\}$ converges to some $x^* \in X^*$ with probability 1.*

Proof The proof of the first part is nearly identical to the corresponding part of Proposition 6. To prove the second part, similar to the proof of Proposition 7, we obtain for all k and all $x^* \in X^*$,

$$E \left\{ \|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k \right\} \leq \|x_k - x^*\|^2 - \frac{2\alpha_k}{m} (F(x_k) - F^*) + \beta \alpha_k^2 c^2 \quad (58)$$

[cf. Eq. (50) with α and y replaced with α_k and x^* , respectively], where $\mathcal{F}_k = \{x_k, z_{k-1}, \dots, z_0, x_0\}$. By the Supermartingale Convergence Theorem (Proposition 2), for each $x^* \in X^*$, we have for all sample paths in a set Ω_{x^*} of probability 1

$$\sum_{k=0}^{\infty} \frac{2\alpha_k}{m} (F(x_k) - F^*) < \infty, \quad (59)$$

and the sequence $\{\|x_k - x^*\|\}$ converges.

Let $\{v_i\}$ be a countable subset of the relative interior $\text{ri}(X^*)$ that is dense in X^* [such a set exists since $\text{ri}(X^*)$ is a relatively open subset of the affine hull of X^* ; an example

of such a set is the intersection of X^* with the set of vectors of the form $x^* + \sum_{i=1}^p r_i \xi_i$, where ξ_1, \dots, ξ_p are basis vectors for the affine hull of X^* and r_i are rational numbers]. The intersection $\bar{\Omega} = \cap_{i=1}^\infty \Omega_{v_i}$ has probability 1, since its complement $\bar{\Omega}^c$ is equal to $\cup_{i=1}^\infty \Omega_{v_i}^c$ and

$$\text{Prob} \left(\cup_{i=1}^\infty \Omega_{v_i}^c \right) \leq \sum_{i=1}^\infty \text{Prob} \left(\Omega_{v_i}^c \right) = 0.$$

For each sample path in $\bar{\Omega}$, all the sequences $\{\|x_k - v_i\|\}$ converge so that $\{x_k\}$ is bounded, while by the first part of the proposition [or Eq. (59)] $\liminf_{k \rightarrow \infty} F(x_k) = F^*$. Therefore, $\{x_k\}$ has a limit point \bar{x} in X^* . Since $\{v_i\}$ is dense in X^* , for every $\epsilon > 0$ there exists $v_{i(\epsilon)}$ such that $\|\bar{x} - v_{i(\epsilon)}\| < \epsilon$. Since the sequence $\{\|x_k - v_{i(\epsilon)}\|\}$ converges and \bar{x} is a limit point of $\{x_k\}$, we have $\lim_{k \rightarrow \infty} \|x_k - v_{i(\epsilon)}\| < \epsilon$, so that

$$\limsup_{k \rightarrow \infty} \|x_k - \bar{x}\| \leq \lim_{k \rightarrow \infty} \|x_k - v_{i(\epsilon)}\| + \|v_{i(\epsilon)} - \bar{x}\| < 2\epsilon.$$

By taking $\epsilon \rightarrow 0$, it follows that $x_k \rightarrow \bar{x}$. □

5 Applications

In this section we illustrate our methods in the context of two types of practical applications, and discuss relations with known algorithms.

5.1 Regularized least squares

Many problems in statistical inference, machine learning, and signal processing involve minimization of a sum of component functions $f_i(x)$ that correspond to errors between data and the output of a model that is parameterized by a vector x . A classical example is least squares problems, where f_i is quadratic. Often a convex regularization function $R(x)$ is added to the least squares objective, to induce desirable properties of the solution. This gives rise to problems of the form

$$\begin{aligned} &\text{minimize} && R(x) + \frac{1}{2} \sum_{i=1}^m (c_i x - d_i)^2 \\ &\text{subject to} && x \in \mathfrak{N}^n, \end{aligned} \tag{60}$$

where c_i and d_i are given vectors and scalars, respectively. When R is differentiable, and either m is very large or the data (c_i, d_i) become available sequentially over time, it makes sense to consider incremental gradient methods, which have a long history of applications over the last 50 years, starting with the Widrow–Hoff least mean squares (LMS) method [58].

The classical type of regularization involves a quadratic function R (as in classical regression and the LMS method), but nondifferentiable regularization functions have become increasingly important recently. On the other hand, to apply our incremental methods, a quadratic R is not essential. What is important is that R has a simple form that facilitates the use of proximal algorithms, such as for example a separable form, so that the proximal iteration on R is simplified through decomposition. As an example, consider the ℓ_1 -regularization problem, where

$$R(x) = \gamma \|x\|_1 = \gamma \sum_{j=1}^n |x^j|, \tag{61}$$

γ is a positive scalar, and x^j is the j th coordinate of x . Then the proximal iteration

$$z_k = \arg \min_{x \in \Re^n} \left\{ \gamma \|x\|_1 + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

decomposes into the n scalar minimizations

$$z_k^j = \arg \min_{x^j \in \Re} \left\{ \gamma |x^j| + \frac{1}{2\alpha_k} |x^j - x_k^j|^2 \right\}, \quad j = 1, \dots, n,$$

and can be done in closed form

$$z_k^j = \begin{cases} x_k^j - \gamma\alpha_k & \text{if } \gamma\alpha_k \leq x_k^j, \\ x_k^j & \text{if } -\gamma\alpha_k < x_k^j < \gamma\alpha_k, \\ x_k^j + \gamma\alpha_k & \text{if } x_k^j \leq -\gamma\alpha_k, \end{cases} \quad j = 1, \dots, n. \tag{62}$$

We refer to Figueiredo et al. [24,57], Beck and Teboulle [10], and the references given there, for a discussion of a broad variety of applications in estimation and signal processing problems, where nondifferentiable regularization functions play an important role.

We now note that the incremental algorithms of this paper are well-suited for solution of ℓ_1 -regularization problems of the form (60)–(61). For example, the k th incremental iteration may consist of selecting a data pair (c_{i_k}, d_{i_k}) and performing a proximal iteration of the form (62) to obtain z_k , followed by a gradient iteration on the component $\frac{1}{2} (c'_{i_k}x - d_{i_k})^2$, starting at z_k :

$$x_{k+1} = z_k - \alpha_k c_{i_k} (c'_{i_k} z_k - d_{i_k}).$$

This algorithm is the special case of the algorithms (19)–(21) (here $X = \Re^n$, and all three algorithms coincide), with $f_i(x)$ being $\gamma \|x\|_1$ (we use m copies of this function) and $h_i(x) = \frac{1}{2} (c'_i x - d_i)^2$. It can be viewed as an incremental version of a popular class of algorithms in signal processing, known as iterative shrinkage/thresholding

(see Chambolle et al. [18], Figueiredo and Nowak [23], Daubechies, et al. [21], Combettes and Wajs [20], Elad et al. [22], Bioucas-Dias and Figueiredo [17], Vonesch and Unser [56], Beck and Teboulle [9,10]). Our methods bear the same relation to this class of algorithms as the LMS method bears to gradient algorithms for the classical linear least squares problem with quadratic regularization function.

Finally, let us note that as an alternative, the proximal iteration (62) could be replaced by a proximal iteration on $\gamma |x^j|$ for some selected index j , with all indexes selected cyclically in incremental iterations. Randomized selection of the data pair (c_{i_k}, d_{i_k}) would also be interesting, particularly in contexts where the data has a natural stochastic interpretation.

5.2 Iterated projection algorithms

A feasibility problem that arises in many contexts involves finding a point with certain properties within a set intersection $\cap_{i=1}^m X_i$, where each X_i is a closed convex set. For the case where m is large and each of the sets X_i has a simple form, incremental methods that make successive projections on the component sets X_i have a long history (see e.g., Gubin et al. [25], and recent papers such as Bauschke [6], Bauschke et al. [2,3], and Cegielski and Suchocka [19], and their bibliographies). We may consider the following generalized version of the classical feasibility problem,

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \cap_{i=1}^m X_i, \end{aligned} \tag{63}$$

where $f : \mathfrak{R}^n \mapsto \mathfrak{R}$ is a convex cost function, and the method

$$x_{k+1} = P_{X_{i_k}} \left(x_k - \alpha_k \tilde{\nabla} f(x_k) \right), \tag{64}$$

where the index i_k is chosen from $\{1, \dots, m\}$ according to a randomized rule. The incremental approach is particularly well-suited for problems of the form (63) where the sets X_i are not known in advance, but are revealed as the algorithm progresses. We note that incremental algorithms for problem (63), which bear some relation with ours have been recently proposed by Nedić [42]. Actually, the algorithm of [42] involves an additional projection on a special set X_0 at each iteration, but for simplicity we take $X_0 = \mathfrak{R}^n$.

While the problem (63) does not involve a sum of component functions, it may be converted into one that does by using an exact penalty function. In particular, consider the problem

$$\begin{aligned} &\text{minimize} && f(x) + \gamma \sum_{i=1}^m \text{dist}(x; X_i) \\ &\text{subject to} && x \in \mathfrak{R}^n, \end{aligned} \tag{65}$$

where γ is a positive penalty parameter. Then for f Lipschitz continuous and γ sufficiently large, problems (63) and (65) are equivalent. We show this for the case where $m = 1$ and then we generalize.

Proposition 10 *Let $f : Y \mapsto \Re$ be a function defined on a subset Y of \Re^n , and let X be a nonempty closed subset of Y . Assume that f is Lipschitz continuous over Y with constant L , i.e.,*

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in Y,$$

and let γ be a scalar with $\gamma > L$. Then the set of minima of f over X coincides with the set of minima of

$$f(x) + \gamma \operatorname{dist}(x; X)$$

over Y .

Proof Denote $F(x) = f(x) + \gamma \operatorname{dist}(x; X)$. For a vector $x \in Y$, let \hat{x} denote a vector of X that is at minimum distance from x . If $\gamma > L$, we have using the Lipschitz property of f ,

$$\begin{aligned} F(x) &= f(x) + \gamma\|x - \hat{x}\| = f(\hat{x}) + (f(x) - f(\hat{x})) + \gamma\|x - \hat{x}\| \\ &\geq f(\hat{x}) = F(\hat{x}), \quad \forall x \in Y, \end{aligned}$$

with strict inequality if $x \neq \hat{x}$. Hence the minima of F over Y can only lie within X , while $F = f$ within X . This shows that if $\gamma > L$, then x^* minimizes f over X if and only if x^* minimizes F over Y . □

We now provide a generalization for $m > 1$.

Proposition 11 *Let $f : Y \mapsto \Re$ be a function defined on a subset Y of \Re^n , and let $X_i, i = 1, \dots, m$, be closed subsets of Y with nonempty intersection. Assume that f is Lipschitz continuous over Y . Then there is a scalar $\bar{\gamma} > 0$ such that for all $\gamma \geq \bar{\gamma}$, the set of minima of f over $\bigcap_{i=1}^m X_i$ coincides with the set of minima of*

$$f(x) + \gamma \sum_{i=1}^m \operatorname{dist}(x; X_i)$$

over Y .

Proof For positive scalars $\gamma_1, \dots, \gamma_m$, and $k = 1, \dots, m$, define

$$F^k(x) = f(x) + \gamma_1 \operatorname{dist}(x; X_1) + \dots + \gamma_k \operatorname{dist}(x; X_k),$$

and for $k = 0$, denote $F^0(x) = f(x)$, $\gamma_0 = 0$. Let L denote the Lipschitz constant for f . By applying Proposition 10, the set of minima of F^m over Y coincides with the

set of minima of F^{m-1} over X_m provided γ_m is greater than $L + \gamma_1 + \dots + \gamma_{m-1}$, the Lipschitz constant for F^{m-1} . Similarly, we obtain that for all $k = 1, \dots, m$, the set of minima of F^k over $\cap_{i=k+1}^m X_i$ coincides with the set of minima of F^{k-1} over $\cap_{i=k}^m X_i$, provided $\gamma_k > L + \gamma_1 + \dots + \gamma_{k-1}$. Thus, the set of minima of F^m over Y coincides with the set of minima of f over $\cap_{i=1}^m X_i$, provided the scalars $\gamma_1, \dots, \gamma_m$ satisfy

$$\gamma_k > L + \gamma_1 + \dots + \gamma_{k-1}, \quad \forall k = 1, \dots, m,$$

where $\gamma_0 = 0$. For such $\gamma_1, \dots, \gamma_m$, the set of minima of $f + \gamma \sum_{i=1}^m \text{dist}(\cdot; X_i)$ over Y coincides with the set of minima of F^m over Y if $\gamma \geq \gamma_m$, and hence also with the set of minima of f over $\cap_{i=1}^m X_i$. □

Note that while the penalty parameter thresholds derived in the preceding proof are quite large, lower thresholds may hold under additional assumptions, such as for convex f and polyhedral X_i . Regarding algorithmic solution, from Proposition 11, it follows that we may consider in place of the original problem (63) the additive cost problem (65) for which our algorithms apply. In particular, let us consider the algorithms (19)–(21), with $X = \mathfrak{N}^n$, which involve a proximal iteration on one of the functions $\gamma \text{dist}(x; X_i)$ followed by a subgradient iteration on f . A key fact here is that the proximal iteration

$$z_k = \arg \min_{x \in \mathfrak{N}^n} \left\{ \gamma \text{dist}(x; X_{i_k}) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\} \tag{66}$$

involves a projection on X_{i_k} of x_k , followed by an interpolation. This is shown in the following proposition.

Proposition 12 *Let z_k be the vector produced by the proximal iteration (66). If $x_k \in X_{i_k}$ then $z_k = x_k$, while if $x_k \notin X_{i_k}$,*

$$z_k = \begin{cases} (1 - \beta_k)x_k + \beta_k P_{X_{i_k}}(x_k) & \text{if } \beta_k < 1, \\ P_{X_{i_k}}(x_k) & \text{if } \beta_k \geq 1, \end{cases} \tag{67}$$

where

$$\beta_k = \frac{\alpha_k \gamma}{\text{dist}(x_k; X_{i_k})}.$$

Proof The case $x_k \in X_{i_k}$ is evident, so assume that $x_k \notin X_{i_k}$. From the nature of the cost function in Eq. (66) we see that z_k is a vector that lies in the line segment between x_k and $P_{X_{i_k}}(x_k)$. Hence there are two possibilities: either

$$z_k = P_{X_{i_k}}(x_k), \tag{68}$$

or $z_k \notin X_{i_k}$ in which case by setting to 0 the gradient at z_k of the cost function in Eq. (66) yields

$$\gamma \frac{z_k - P_{X_{i_k}}(z_k)}{\|z_k - P_{X_{i_k}}(z_k)\|} = \frac{1}{\alpha_k}(x_k - z_k).$$

Hence $x_k, z_k,$ and $P_{X_{i_k}}(z_k)$ lie on the same line, so $P_{X_{i_k}}(z_k) = P_{X_{i_k}}(x_k)$ and

$$z_k = x_k - \frac{\alpha_k \gamma}{\text{dist}(x_k; X_{i_k})} (x_k - P_{X_{i_k}}(x_k)) = (1 - \beta_k)x_k + \beta_k P_{X_{i_k}}(x_k). \tag{69}$$

By calculating and comparing the value of the cost function in Eqs. (66) for each of the possibilities (68) and (69), we can verify that (69) gives a lower cost if and only if $\beta_k < 1$. □

Let us finally note that our incremental methods also apply to the problem

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^m f_i(x) \\ &\text{subject to} && x \in \cap_{i=1}^m X_i. \end{aligned}$$

In this case the interpolated projection iterations (67) on the sets X_i are followed by subgradient or proximal iterations on the components f_i . A related problem is

$$\begin{aligned} &\text{minimize} && f(x) + c \sum_{j=1}^r \max\{0, g_j(x)\} \\ &\text{subject to} && x \in \cap_{i=1}^m X_i, \end{aligned}$$

which is obtained by replacing convex inequality constraints of the form $g_j(x) \leq 0$ with the nondifferentiable penalty terms $c \max\{0, g_j(x)\}$, where $c > 0$ is a penalty parameter. Then a possible incremental method at each iteration, would either do a subgradient iteration on f , or select one of the violated constraints (if any) and perform a subgradient iteration on the corresponding function g_j , or select one of the sets X_i and do an interpolated projection on it. Except for the projections on X_i , variants of this algorithm are well-known.

6 Conclusions

The incremental proximal algorithms of this paper provide new possibilities for minimization of many-term sums of convex component functions. It is generally believed that proximal iterations are more stable than gradient and subgradient iterations. It may thus be important to have flexibility to separate the cost function into the parts that are conveniently handled by proximal iterations (e.g., in essentially closed form), and the

remaining parts to be handled by subgradient iterations. We provided a convergence analysis and showed that our algorithms are well-suited for some problems that have been the focus of recent research.

Much work remains to be done to apply and evaluate our methods within the broad context of potential applications. Let us mention some possibilities that may extend the range of applications of our approach, and are interesting subjects for further investigation: alternative proximal and projected subgradient iterations, involving nonquadratic proximal terms and/or subgradient projections, alternative stepsize rules, distributed asynchronous implementations along the lines of [38], polyhedral approximation (bundle) variants of the proximal iterations in the spirit of [11], and variants for methods with errors in the calculation of the subgradients along the lines of [41].

References

1. Blatt, D., Hero, A.O., Gauchman, H.: A convergent incremental gradient method with a constant step size. *SIAM J. Optim.* **18**, 29–51 (2008)
2. Bauschke, H.H., Combettes, P.L., Luke, D.R.: Hybrid projection-reflection method for phase retrieval. *J. Opt. Soc. Am.* **20**, 1025–1034 (2003)
3. Bauschke, H.H., Combettes, P.L., Kruk, S.G.: Extrapolation algorithm for affine-convex feasibility problems. *Numer. Algorithms* **41**, 239–274 (2006)
4. Ben-Tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization method and its use for the positron emission tomography reconstruction. In: Butnariu, D., Censor, Y., Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Elsevier, Amsterdam, Netherlands (2001)
5. Bertsekas, D.P., Nedić, C.A., Ozdaglar, A.E.: *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA (2003)
6. Bauschke, H.H.: Projection algorithms: results and open problems. In: Butnariu, D., Censor, Y., Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Elsevier, Amsterdam, Netherlands (2001)
7. Bertsekas, D.P., Tsitsiklis, J.N.: *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA (1996)
8. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods. *SIAM J. Optim.* **10**, 627–642 (2000)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2**, 183–202 (2009)
10. Beck, A., Teboulle, M.: Gradient-based algorithms with applications to signal-recovery problems. In: Eldar, Y., Palomar, D. (eds.) *Convex Optimization in Signal Processing and Communications*, pp. 42–88. Cambridge University Press, Cambridge (2010)
11. Bertsekas, D.P., Yu, H.: A unifying polyhedral approximation framework for convex optimization. In: *Laboratory for Information and Decision Systems Report LIDS-P-2820*. MIT (2009); *SIAM J. Optim.* (to appear)
12. Bertsekas, D.P.: Incremental least squares methods and the extended Kalman filter. *SIAM J. Optim.* **6**, 807–822 (1996)
13. Bertsekas, D.P.: Hybrid incremental gradient method for least squares. *SIAM J. Optim.* **7**, 913–926 (1997)
14. Bertsekas, D.P.: *Nonlinear Programming*. 2nd edn. Athena Scientific, Belmont, MA (1999)
15. Bertsekas, D.P.: *Convex Optimization Theory*. Athena Scientific, Belmont, MA (2009)
16. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. In: *Laboratory for Information and Decision Systems Report LIDS-P-2848*. MIT (2010)
17. Bioucas-Dias, J., Figueiredo, M.A.T.: A new TwIST: two-step iterative shrinkage thresholding algorithms for image restoration. *IEEE Trans. Image Process.* **16**, 2992–3004 (2007)

18. Chambolle, A., DeVore, R.A., Lee, N.Y., Lucier, B.J.: Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.* **7**, 319–335 (1998)
19. Cegielski, A., Suchocka, A.: Relaxed alternating projection methods. *SIAM J. Optim.* **19**, 1093–1106 (2008)
20. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.* **4**, 1168–1200 (2005)
21. Daubechies, I., Defrise, M., Mol, C.D.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.* **57**, 1413–1457 (2004)
22. Elad, M., Matalon, B., Zibulevsky, M.: Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *J. Appl. Comput. Harmon. Anal.* **23**, 346–367 (2007)
23. Figueiredo, M.A.T., Nowak, R.D.: An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.* **12**, 906–916 (2003)
24. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.* **1**, 586–597 (2007)
25. Gubin, L.G., Polyak, B.T., Raik, E.V.: The method of projection for finding the common point in convex sets. *U.S.S.R. Comput. Math. Phys.* **7**, 1–24 (English Translation) (1967)
26. Grippo, L.: A class of unconstrained minimization methods for neural network training. *Optim. Methods Softw.* **4**, 135–150 (1994)
27. Grippo, L.: Convergent on-line algorithms for supervised learning in neural networks. *IEEE Trans. Neural Netw.* **11**, 1284–1299 (2000)
28. Helou, E.S., De Pierro, A.R.: Incremental subgradients for constrained convex optimization: a unified framework and new methods. *SIAM J. Optim.* **20**, 1547–1572 (2009)
29. Johansson, B., Rabi, M., Johansson, M.: A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM J. Optim.* **20**, 1157–1170 (2009)
30. Kibardin, V.M.: Decomposition into functions in the minimization problem. *Autom. Remote Control* **40**, 1311–1323 (1980)
31. Kiwiel, K.C.: Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM J. Optim.* **14**, 807–840 (2004)
32. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* **16**, 964–979 (1979)
33. Litvakov, B.M.: On an iteration method in the problem of approximating a function from a finite number of observations. *Avtom. Telemekh.* **4**, 104–113 (1966)
34. Luo, Z.Q., Tseng, P.: Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. *Optim. Methods Softw.* **4**, 85–101 (1994)
35. Luo, Z.Q.: On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks. *Neural Comput.* **3**, 226–245 (1991)
36. Mangasarian, O.L., Solodov, M.V.: Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optim. Methods Softw.* **4**, 103–116 (1994)
37. Martinet, B.: Regularisation d’in équations variationnelles par approximations successives. *Revue Fran. d’Automatique Et Infomatique Rech. Op’ Erationelle* **4**, 154–159 (1970)
38. Nedić, A., Bertsekas, D.P., Borkar, V.: Distributed asynchronous incremental subgradient methods. In: Butnariu, D., Censor, Y., Reich, S. (eds.) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Elsevier, Amsterdam, Netherlands (2001)
39. Nedić, A., Bertsekas, D.P.: Convergence rate of the incremental subgradient algorithm. In: Uryasev, S., Pardalos, P.M. (eds.) *Stochastic Optimization: Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht (2000)
40. Nedić, A., Bertsekas, D.P.: Incremental subgradient methods for nondifferentiable optimization. *SIAM J. Optim.* **12**, 109–138 (2001)
41. Nedić, A., Bertsekas, D.P.: The effect of deterministic noise in subgradient methods. *Math. Program. Ser. A* **125**, 75–99 (2010)
42. Nedić, A.: Random projection algorithms for convex minimization problems. University of Illinois Report (2010); *Math. Program. J.* (to appear)
43. Neveu, J.: *Discrete Parameter Martingales*. North-Holland, Amsterdam, The Netherlands (1975)
44. Predd, J.B., Kulkarni, S.R., Poor, H.V.: A collaborative training algorithm for distributed learning. *IEEE Trans. Inf. Theory* **55**, 1856–1871 (2009)

45. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.* **72**, 383–390 (1979)
46. Ram, S.S., Nedić, A., Veeravalli, V.V.: Incremental stochastic subgradient algorithms for convex optimization. *SIAM J. Optim.* **20**, 691–717 (2009)
47. Ram, S.S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. *J. Optim. Theory Appl.* **147**, 516–545 (2010)
48. Rabbat, M.G., Nowak, R.D.: Distributed optimization in sensor networks. In: *Proceedings of Information Processing Sensor Networks*, pp. 20–27. Berkeley, CA (2004)
49. Rabbat, M.G., Nowak, R.D.: Quantized incremental algorithms for distributed optimization. *IEEE J. Sel. Areas Commun.* **23**, 798–808 (2005)
50. Rockafellar, R.T.: *Convex Analysis*. Princeton University Press, Princeton (1970)
51. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**, 877–898 (1976)
52. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter A.: Pegasos: primal estimated subgradient solver for SVM. In: *ICML 07* pp. 807–814. New York, N.Y. (2007)
53. Solodov, M.V., Zavriev, S.K.: Error stability properties of generalized gradient-type algorithms. *J. Opt. Theory Appl.* **98**, 663–680 (1998)
54. Solodov, M.V.: Incremental gradient algorithms with stepsizes bounded away from zero. *Comput. Optim. Appl.* **11**, 28–35 (1998)
55. Tseng, P.: An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM J. Optim.* **8**, 506–531 (1998)
56. Vonesch, C., Unser, M.: Fast iterative thresholding algorithm for wavelet-regularized deconvolution. In: *Proceedings of the SPIE Optics and Photonics 2007 Conference on Mathematical Methods: Wavelet XII*, vol. 6701, pp. 1–5. San Diego, CA (2007)
57. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 3373–3376 (2008)
58. Widrow, B., Hoff, M.E.: Adaptive switching circuits. Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, 96–104 (1960)