DISTRIBUTED ASYNCHRONOUS INCREMENTAL SUBGRADIENT METHODS *

A. Nedić, D. P. Bertsekas,[a†] and V. S. Borkar[b‡]

[a]Massachusetts Institute of Technology, Rm. 35-210,
77 Massachusetts Ave., Cambridge, MA 02139 USA

[b]School of Technology and Computer Science,
Tata Institute of Fundamental Research,
Homi Bhabha Road, Mumbai 400005, India

We propose and analyze a distributed asynchronous subgradient method for minimizing a convex function that consists of the sum of a large number of component functions. This type of minimization arises in a dual context from Lagrangian relaxation of the coupling constraints of large scale separable problems. The idea is to distribute the computation of the component subgradients among a set of processors, which communicate only with a coordinator. The coordinator performs the subgradient iteration incrementally and asynchronously, by taking steps along the subgradients of the component functions that are available at the update time. The incremental approach has performed very well in centralized computation, and the parallel implementation should improve its performance substantially, particularly for typical problems where computation of the component subgradients is relatively costly.

1. Introduction

We focus on the problem

$$\begin{array}{ll} \text{minimize} & f(x) = f_1(x) + \ldots + f_m(x) \\ \text{subject to} & x \in X, \end{array} \tag{1}$$

where $f_i : \Re^n \to \Re$ are convex functions and X is a nonempty, closed, and convex subset of $\Re^n$. We are primarily concerned with the case where f is nondifferentiable. A special case of particular interest arises in Lagrangian relaxation where f is the dual function of a primal separable combinatorial problem that is solved by decomposition. Within this

*Appeared in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., Eds., Studies in Computational Mathematics, Elsevier, Amsterdam 2001, pp. 381–407.
†Research supported by NSF under Grant ACI-9873339.
‡Research supported by NSF-KDI under Grant ECS9873451.

context, x is a vector of Lagrange multipliers corresponding to coupling constraints, and the component functions $f_i(x)$ are obtained by solving subproblems corresponding to the separable terms in the cost function and the constraints.

In this paper, we consider the solution of problem (1) using distributed asynchronous versions of the classical subgradient iteration

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i=1}^{m} g_i(t) \right]. \tag{2}$$

Here $g_i(t)$ is a subgradient of f_i at $x(t)$, $\alpha(t)$ is a positive stepsize, and $\mathcal{P}_X$ denotes the projection on the set $X \subset \Re^n$.

The most straightforward way to parallelize the above iteration is to use multiple processors to compute in parallel the component subgradients $g_i(t)$. Once all of these components have been computed, they can be collected at a single processor, called the *updating processor*, who will execute the update of the vector $x(t)$ using iteration (2). The updating processor will then distribute/broadcast in some way the new iterate $x(t+1)$ to the subgradient-computing processors who will collectively compute the new subgradients for the subsequent iteration. This parallelization approach is quite efficient as long as the computation of the subgradients (which is parallelized) takes much more time than their addition and the execution of the iteration (2) (which are performed serially). Fortunately, this is typically the case in the principal context of interest to us, i.e., duality and Lagrangian relaxation.

The parallel algorithm just described is mathematically equivalent to the serial iteration (2). It can be termed *synchronous*, in the sense that there is clear division between the computations of successive iterations, i.e., all computation relating to iteration t must be completed before iteration $t+1$ can begin. In this paper we are interested in *asynchronous* versions of the subgradient method (2), where the subgradient components at a given iteration do not necessarily correspond to the same value of x. An example of such a method is

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i=1}^{m} g_i(\tau_i(t)) \right], \tag{3}$$

where for all i, we have $\tau_i(t) \leq t$ and the difference $t - \tau_i(t)$ may be viewed as a "delay." Such an iteration may be useful if for some reason (e.g., excessive computation or communication delay) some subgradient components $g_i(t)$ are not available at time t, and to avoid further delay in executing the update of $x(t)$, the most recently computed components $g_i(\tau_i(t))$ are used in (3) in place of the missing components $g_i(t)$.

A more general version of the asynchronous iteration (3) is given by

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i \in I(t)} g_i(\tau_i(t)) \right], \tag{4}$$

where $I(t)$ is a nonempty subset of the index set $\{1, \ldots, m\}$, $\tau_i(t)$ satisfies $\tau_i(t) \leq t$ for all t and i, $g_i(\tau_i(t))$ is a subgradient of f_i computed at $x(\tau_i(t))$, and $x(0) \in X$ is an initial

point. To visualize the execution of this iteration, it is useful to think of the computing system as consisting of two parts: the *updating system* (US for short), and the *subgradient computing system* (GCS for short) (see Fig. 1).

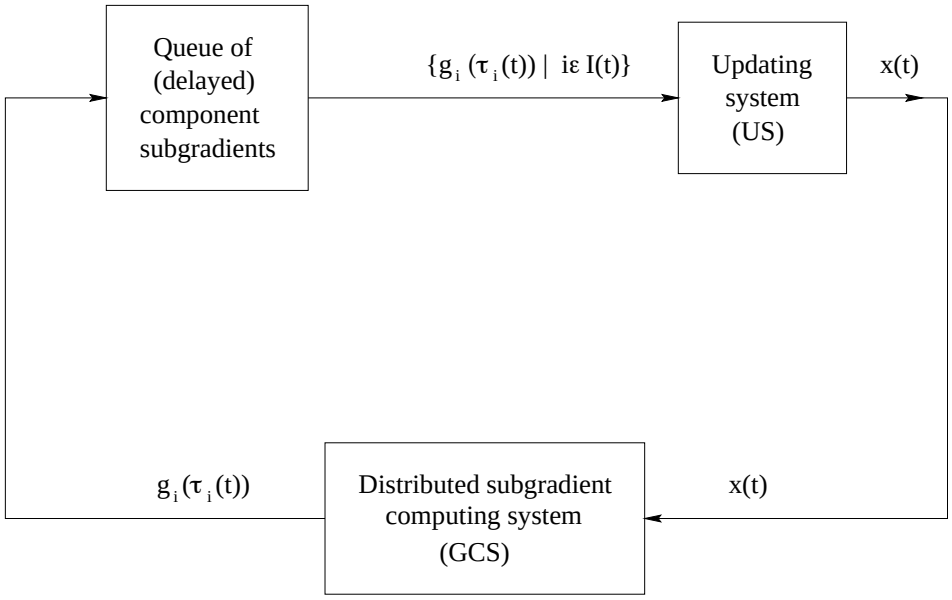

Figure 1. Visualization of the implementation of the asynchronous distributed iteration

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i \in I(t)} g_i(\tau_i(t)) \right].$$

This iteration is executed by the updating system (US), using subgradients deposited in a queue by the distributed subgradient computing system (GCS).

The US executes iteration (4) at each time t and delivers the corresponding iterates $x(t)$ to the GCS. The GCS uses the values $x(t)$ obtained from the US, computes subgradient components $g_i(\tau_i(t))$, and deposits them in a queue from where they can be accessed by the US. There is no synchronization between the operations of the CS and the GCS. Furthermore, while the GCS may involve multiple processors that compute subgradient components in parallel, the characteristics of the GCS (e.g., shared memory, message passing, communication architecture, number of processors, synchronization rules, etc.) are not material to the description of our algorithm.

The motivation for considering iteration (4) rather than its special case (3) is twofold. First, it makes sense to keep the US busy with updates while the GCS is computing subgradient components. This is particularly so if the computation of g_i is much more time consuming for some i than for others, thereby creating a synchronization bottleneck.

Second, it appears that updating the value of $x(t)$ as quickly as possible and using it in the calculation of the component subgradients has a beneficial effect in the convergence rate of the subgradient method. This is the main characteristic of *incremental subgradient*

methods, which we have studied in [19], and which may be viewed as the special case of iteration (4) where $\tau_i(t) = t$ for all i and t, and the set $I(t)$ consists of a *single* index.

We note that incremental methods for differentiable unconstrained problems have a long tradition, most notably in the training of neural networks, where they are known as *backpropagation methods*. They are related to the Widrow-Hoff algorithm [25] and to stochastic gradient/stochastic approximation methods, and they are supported by several recent convergence analyses Bertsekas [3], Bertsekas and Tsitsiklis [5]–[6], Gaivoronski [8], Grippo [10], Luo [16], Luo and Tseng [17], Mangasarian and Solodov [18], Tseng [23]. It has been experimentally observed that incremental gradient methods often converge much faster than the steepest descent method when far from the eventual limit. However, near convergence, they typically converge slowly because they require a diminishing stepsize [e.g., $\alpha_k = O(1/k)$] for convergence. If α_k is instead taken to be a small enough constant, "convergence" to a limit cycle occurs, as first shown by Luo [16].

The incremental subgradient method was studied first by Kibardin [12], and more recently by Solodov and Zavriev [22], Nedić and Bertsekas [19]–[21], and Ben-Tal, Margalit and Nemirovski [1]; the incremental subgradient method of the form (3) is considered by Zhao, Luh, and Wang [26], Kiwiel and Lindberg [13]; the incremental ϵ-subgradient method is analyzed by Kiwiel [14]; applications of incremental methods can be found in Kaskavelis and Caramanis [11], Ben-Tal, Margalit and Nemirovski [1]. As we have discussed in our papers [19] and [21], incremental subgradient methods exhibit a similar behavior to incremental gradient methods. We believe that the incremental structure that is inherent in our proposed parallel subgradient method (4) results in convergence and rate of convergence characteristics that are similar to those of incremental gradient and subgradient methods. In particular, we expect an enhanced convergence rate over the nonincremental version given by Eq. (3).

In this paper, we will analyze the following version of iteration (4), given by

$$x(t+1) = \mathcal{P}_X\left[x(t) - \alpha(t) g_{i(t)}(\tau(t))\right]. \tag{5}$$

Here we assume for simplicity that the set $I(t)$ consists of a single element denoted $i(t)$. For $X = \Re^n$, this simplification does not involve an essential loss of generality since an iteration involving multiple component function subgradients may be broken down into several iterations each involving a single component function subgradient. When $X \neq \Re^n$, our analysis can be extended for the more general iteration (4).

The most important assumptions in our analysis are:

(a) The stepsize $\alpha(t)$ is either constant, or is diminishing to 0 and satisfies some common technical conditions such as $\sum_{t=0}^{\infty} \alpha(t) = \infty$ (see a more precise statement later). In the case of a constant stepsize, we only show convergence to optimality within an error which depends on the length of the stepsize.

(b) The "delay" $t - \tau(t)$ is bounded from above by some (unknown) positive integer D, so that our algorithm belongs to the class of *partially asynchronous methods*, as defined by Bertsekas and Tsitsiklis [4].

(c) All the component functions f_i are used with the same "long-term frequency" by

the algorithm. Precise methods to enforce this assumption are given later, but basically what we mean is that if $n_i(t)$ is the number of times up to iteration t where a subgradient of the component f_i is used by the algorithm, then the ratios $n_i(t)/t$ should all be asymptotically equal to $1/m$ (as $t \to \infty$).

(d) The subgradients $g_{i(t)}(\tau(t))$ used in the method are bounded.

The restriction (c) can be easily enforced in a number of ways, by regulating the frequency of the indices of subgradient components computed by the subgradient computing system. We will consider one specific approach, whereby we first select a sequence of indexes $\{j(t)\}$ according to one of two rules:

(1) The *cyclic rule* where the sequence $\{j(t)\}$ is obtained by a permutation of each of the periodic blocks $\{1, 2, \ldots, m\}$ in the periodic sequence $\{1, 2, \ldots, m, 1, 2, \ldots, m, \ldots\}$.

(2) The *random rule* where $\{j(t)\}$ is a sequence of independent identically distributed random variables, each taking the values $1, 2, \ldots, m$ with equal probability $1/m$.

Given a sequence $\{j(t)\}$ obtained by the cyclic or the random rule, the sequence $\{i(t)\}$ used in iteration (5) is given by

$$i(t) = j(\pi(t)), \tag{6}$$

where $\pi(\cdot)$ is a permutation mapping that maps the set $\{0, 1, \ldots\}$ into itself, such that for some positive integer T, we have

$$|\pi(t) - t| \leq T, \qquad \forall\ t = 0, 1, \ldots. \tag{7}$$

The permutation mapping $\pi(\cdot)$ captures the asynchronous character of the algorithm, whereby component function subgradients are offered to the updating system in the order of $\{j(\pi(t))\}$, which is different than the order of $\{j(t)\}$ in which their computation was initiated within the subgradient computing system.

A version of the algorithm that does not work is when the component subgradients $g_{i(t)}(\tau(t))$ are normalized by multiplying with $1/||g_{i(t)}(\tau(t))||$, which may be viewed as a weight associated with the component $f_{i(t)}$ at time t. Unless these weights are asymptotically equal, this modification would effectively alter the effective "long-term frequency" by which the components f_i are selected, thereby violating a fundamental premise for the validity of our algorithm.

We note that our proposed parallel algorithms (4) and (5) do not fit the framework of the general algorithmic models of Chapters 6 and 7 of Bertsekas and Tsitsiklis [4], so it is not covered by the line of analysis of that reference. In the latter models, at each time t, only *some of the components of* x are updated using an equation that (within our subgradient method context) would depend on *all components* f_i (perhaps with communication delays). By contrast in the present paper, at each time t, *all components of* x are updated using an equation that involves some of the components f_i. While it is possible to consider alternative asynchronous subgradient methods of the type given in Chapters 6 and 7 of Bertsekas and Tsitsiklis [4], we believe that these methods would not be as well suited to typical subgradient optimization problems, which arise in the context of Lagrangian relaxation and duality.

The proof ideas of the present paper are related to those of parallel asynchronous deterministic and stochastic gradient methods as discussed in Tsitsiklis, Bertsekas, and Athans [24], and Bertsekas and Tsitsiklis [4], as well as the proof ideas of incremental deterministic and randomized subgradient methods as discussed in Nedić and Bertsekas [19]. In particular, the key proof idea is to view the parallel asynchronous method as an iterative method with deterministic or stochastic errors, the effects of which are controlled with an appropriate mechanism, such as stepsize selection. An alternative approach is possible based on differential inclusions that extend the "ODE" approach for the analysis of stochastic approximation algorithms (see Benveniste, Metivier, and Priouret [2], Borkar [7], and Kushner and Yin [15]).

We note also that while a (nonparallel) incremental subgradient method with a diminishing stepsize can be viewed as an ϵ-subgradient method (under some boundedness assumptions) [20], for the algorithm of this paper, this connection is much harder to make and has not been attempted. The reason is that with delays in the calculations of the component subgradients, it is difficult to estimate the ϵ parameter corresponding to the ϵ-subgradient method. Furthermore, for a different stepsize rule (e.g., a constant stepsize), it is not possible to make any fruitful connection with an ϵ-subgradient method.

The paper is structured as follows. In Section 2, we state and discuss our results for the cyclic rule. In Section 3, we do the same for the random selection rule. In Sections 4 and 5, we give the proofs of the convergence results presented in Sections 2 and 3, respectively.

2. Convergence Results for Cyclic Selection Rule

Throughout the paper, we use f^* and X^* to denote the optimal function value and the optimal solution set for problem (1), respectively. In this section, we present convergence results for the method with the cyclic selection rule, under the following assumption.

Assumption 2.1:

(a) There exists a positive constant C such that for all $i = 1, \ldots, m$

$$||g|| \leq C, \qquad \forall\ g \in \partial f_i(x(\tau(t))) \cup \partial f_i(x(t)), \qquad t = 0, 1 \ldots,$$

where $\partial f_i(x)$ denotes the set of subgradients of f_i at a point x.

(b) There exists a positive integer D such that

$$t - \tau(t) \leq D, \qquad \forall\ t = 0, 1, \ldots.$$

Note that if the components f_i are polyhedral or if the set X is compact, then Assumption 2.1(a) holds. Assumption 2.1(b) is natural, since our algorithm does not use the value of the bound D.

We now give the convergence result for a constant stepsize.

Proposition 2.1: Let Assumption 2.1 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule and the stepsize fixed to some positive scalar α, the following hold:

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

$$\liminf_{t\to\infty} f(x(t)) = -\infty.$$

(b) If f^* is finite, then

$$\liminf_{t\to\infty} f(x(t)) \leq f^* + mC^2\left(\frac{1}{2} + m + 2D + T\right)\alpha.$$

Next, we consider a diminishing stepsize that satisfies the following.

Assumption 2.2: The stepsize $\alpha(t)$ is given by

$$\alpha(t) = \frac{r_0}{(l+r_1)^q}, \qquad t = \sigma_l, \sigma_l+1, \ldots, \sigma_{l+1}-1, \quad l = 0, 1\ldots,$$

where r_0, r_1, and q are some positive scalars with $0 < q \leq 1$, and the sequence $\{\sigma_l\}$ is increasing and is such that $\sigma_{l+1} - \sigma_l \leq S$ for some positive integer S and all l.

For this stepsize we have the following convergence result.

Proposition 2.2: Let Assumptions 2.1 and 2.2 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule, we have

$$\liminf_{t\to\infty} f(x(t)) = f^*.$$

The result of Prop. 2.2 can be strengthened in the case where the optimal solution set X^* is nonempty, under a mild additional restriction on the stepsize. This stronger convergence result is given in the next proposition.

Proposition 2.3: Let Assumptions 2.1 and 2.2 hold, where $\frac{1}{2} < q \leq 1$ in Assumption 2.2. Also, let X^* be nonempty. Then the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule converges to an optimal solution.

3. Convergence Results for a Random Selection Rule

In this section, we present convergence results for the method with the random selection rule. The initial point $x(0) \in X$ and the stepsize $\alpha(t)$ are deterministic. In our analysis we use the following assumption.

Assumption 3.1:

(a) Assumption 2.1 holds.

(b) Assumption 2.2 holds with the scalar q satisfying $\frac{3}{4} < q \leq 1$.

(c) The sequence $\{j(t)\}$ is a sequence of independent random variables each of which is uniformly distributed over the set $\{1, \ldots, m\}$. Furthermore, the sequence $\{j(t)\}$ is independent of the sequence $\{x(t)\}$.

We now give the convergence result for a diminishing stepsize.

Proposition 3.1: Let Assumption 3.1 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the random selection rule, we have with probability 1

$$\liminf_{t \to \infty} f(x(t)) = f^*.$$

When the optimal solution set X^* is nonempty, we can strengthen Prop. 3.1, as shown in the following proposition.

Proposition 3.2: Let Assumption 3.1 hold and let X^* be nonempty. Then the sequence $\{x(t)\}$ generated by the method with the random selection rule converges to some optimal solution with probability 1.

Remark 3.1: If the underlying set X is compact, then it can be shown that Prop. 3.2 holds using a wider range of values for q in the stepsize rule [cf. Assumption 3.1(b)]. In particular, for $\frac{1}{2} < q \leq 1$, the result of Prop. 3.2 is valid, which we discuss in more detail in Section 5 (cf. Remark 5.1).

4. Convergence Proofs for Cyclic Selection Rule

In this section and the next one, for notational convenience, we define $\alpha(t) = \alpha(0)$ for $t < 0$, and $t_k = km$ and $x_k = x(t_k)$ for all k. Here we give the proofs for the convergence results of Section 2. The proofs are complicated, so we break them down in several steps. We first provide some estimates of the progress of the method in terms of the distances of the iterates to an arbitrary point in X and in terms of the objective function values. These estimates are given in the subsequent Lemma 4.2. Some preliminary results needed for the proof of Lemma 4.2 and Lemma 5.2 of Section 5 are given in the following lemma.

Lemma 4.1: Let Assumption 2.1 hold. Then, for the iterates $x(t)$ generated by the method, the following hold:

(a) For all $y \in X$ and t, we have

$$\begin{aligned} ||x(t+1) - y||^2 \;\leq\; & ||x(t) - y||^2 - 2\alpha(t)\Big(f_{j(t)}(x(t)) - f_{j(t)}(y)\Big) \\ & + C^2(1 + 4D)\alpha^2(t - D) \\ & + 2\alpha(t) \sum_{l=1}^{m} \Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big), \end{aligned} \tag{8}$$

where δ^l_i is the Kronecker symbol (i.e., $\delta^l_i = 1$ if $l = i$ and $\delta^l_i = 0$ otherwise).

(b) For all $y \in X$, and N and K with $N \geq K$, we have

$$\sum_{l=1}^{m}\sum_{t=K}^{N}\alpha(t)\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l(x(t))-f_l(y)\Big) \leq C^2T\sum_{t=K}^{N}\alpha^2(t-T)$$

$$+\max\{C,G(y)\}\sum_{t=K}^{N}\Big(\alpha(t-T)-\alpha(t+T)\Big)\left(C\sum_{r=0}^{t}\alpha(r)+||x_0-y||\right)$$

$$+c(y)\Big(\alpha^2(K)+\alpha(K)+\alpha^2(N+1-T)+\beta\alpha(N+1-T)\Big)$$

$$+\left(\alpha(K)||x(K)-y||^2+\frac{1}{\beta}\alpha(N+1-T)||x(N+1)-y||^2\right), \tag{9}$$

where β is an arbitrary positive scalar, and

$$G(y)=\max\{||g|| \mid g\in\partial f_l(y),\ l=1,\ldots,m\}, \tag{10}$$

$$c(y)=\max\left\{CT^2\Big(C+G(y)\Big),\frac{T^2}{2}\Big(C^2+G^2(y)\Big)\right\}. \tag{11}$$

Proof: (a) From the definition of $x(t+1)$ [cf. Eq. (5)], the nonexpansion property of the projection, and the subgradient boundedness [cf. Assumption 2.1(a)], we have for all $y \in X$ and t

$$\begin{aligned}||x(t+1)-y||^2 &\leq ||x(t)-y||^2-2\alpha(t)g_{i(t)}(\tau(t))'\Big(x(t)-y\Big)+C^2\alpha^2(t)\\ &\leq ||x(t)-y||^2-2\alpha(t)\Big(f_{i(t)}(x(t))-f_{i(t)}(y)\Big)\\ &\quad +4C\alpha(t)||x(t)-x(\tau(t))||+C^2\alpha^2(t),\end{aligned} \tag{12}$$

where in the last inequality we use

$$g_{i(t)}(\tau(t))'\Big(x(t)-y\Big)\geq f_{i(t)}(x(t))-f_{i(t)}(y)-2C||x(t)-x(\tau(t))||,$$

which can be obtained by using the fact $x(t)=x(\tau(t))+(x(t)-x(\tau(t)))$, the convexity of $f_{i(t)}$, and the subgradient boundedness. Furthermore, from the relation

$$||x(t)-x(\hat{t})||\leq C\sum_{s=\hat{t}}^{t-1}\alpha(s), \qquad \forall\ t,\hat{t},\ t\geq\hat{t}, \tag{13}$$

and the facts $t-D\leq\tau(t)\leq t$ for all t, and $\alpha(r)\leq\alpha(t-D)$ for $r=t-D,\ldots,t-1$ and all t, we obtain

$$||x(t)-x(\tau(t))||\leq C\sum_{r=t-D}^{t-1}\alpha(r)\leq CD\alpha(t-D).$$

By using this estimate and the fact $\alpha(t) \leq \alpha(t-D)$ for all t in Eq. (12), we have

$$||x(t+1)-y||^2 \leq ||x(t)-y||^2 - 2\alpha(t)\Big(f_{i(t)}(x(t)) - f_{i(t)}(y)\Big) + C^2(1+4D)\alpha^2(t-D),$$

from which, by adding and subtracting $2\alpha(t)\Big(f_{j(t)}(x(t)) - f_{j(t)}(y)\Big)$, and by using the Kronecker symbol, we obtain Eq. (8).

(b) We introduce the following sets:

$$M_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid j(t) = i(p(t)) \text{ with } p(t) \in \{K,\ldots,N\}\Big\},$$

$$P_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid j(t) = i(p(t)) \text{ with } p(t) < K \text{ or } p(t) > N\Big\}, \tag{14}$$

$$Q_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid i(t) = j(\pi(t)) \text{ with } \pi(t) < K \text{ or } \pi(t) > N\Big\}, \tag{15}$$

where $p(t)$ is the inverse of the permutation mapping $\pi(t)$, i.e., $p(t) = \pi^{-1}(t)$. Note that, since $|\pi(t) - t| \leq T$ for all t [cf. Eq. (7)], for the inverse mapping $p(t)$ we have

$$|p(t) - t| \leq T, \qquad \forall\ t = 0, 1, \ldots.$$

The set $M_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{j(t)}$ of $f_{j(t)}$ is used in an update of $x(t)$ at some time between K and N. Similarly, the set $P_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{j(t)}$ of $f_{j(t)}$ is used in an update $x(t)$ at some time before K or after N [i.e., $j(t) = i(p(t)) \notin \{i(K),\ldots,i(N)\}$]. The set $Q_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{i(t)}$ of $f_{i(t)}$ is used in an update $x(t)$ at some time between K and N, but the $j(\pi(t))$ corresponding to $i(t)$ does not belong to the set $\{j(K),\ldots,j(N)\}$. By using the above defined sets, we have

$$\sum_{l=1}^{m}\sum_{t=K}^{N} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big) =$$

$$\sum_{l=1}^{m}\sum_{t\in M_{K,N}} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big)$$

$$+\sum_{l=1}^{m}\sum_{t\in P_{K,N}} \alpha(t)\delta^l_{j(t)}\Big(f_l(x(t)) - f_l(y)\Big)$$

$$-\sum_{l=1}^{m}\sum_{t\in Q_{K,N}} \alpha(t)\delta^l_{i(t)}\Big(f_l(x(t)) - f_l(y)\Big). \tag{16}$$

Next we estimate each of the terms in the preceding relation. According to the definition of $M_{K,N}$, we have $j(t) = i(p(t))$ for all $t \in M_{K,N}$ [i.e., $g_{j(t)}$ is used at time $p(t)$ with the corresponding step $\alpha(p(t))$], so that

$$\sum_{l=1}^{m}\sum_{t\in M_{K,N}} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big) =$$

$$\sum_{l=1}^{m}\sum_{t\in M_{K,N}}\delta^l_{j(t)}\Big[\alpha(t)\Big(f_l(x(t))-f_l(y)\Big)-\alpha(p(t))\Big(f_l(x(p(t)))-f_l(y)\Big)\Big]$$

$$=\sum_{l=1}^{m}\sum_{t\in M_{K,N}}\delta^l_{j(t)}\alpha(p(t))\Big(f_l(x(t))-f_l(x(p(t)))\Big)$$

$$+\sum_{l=1}^{m}\sum_{t\in M_{K,N}}\delta^l_{j(t)}\Big(\alpha(t)-\alpha(p(t))\Big)\Big(f_l(x(t))-f_l(y)\Big).$$

By using the convexity of each f_l, the subgradient boundedness, the monotonicity of $\alpha(t)$, and the facts $|p(t)-t|\le T$ and $\sum_{l=1}^m\delta^l_{j(t)}=1$ for all t, from the preceding relation we obtain

$$\sum_{l=1}^{m}\sum_{t\in M_{K,N}}\alpha(t)\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l(x(t))-f_l(y)\Big)\le C\sum_{t=K}^{N}\alpha(t-T)||x(t)-x(p(t))||$$

$$+\max\{C,G(y)\}\sum_{t=K}^{N}\Big(\alpha(t-T)-\alpha(t+T)\Big)||x(t)-y||,$$

where $G(y)$ is given by Eq. (10). Furthermore, we have

$$||x(t)-x(p(t))||\le CT\alpha(t-T),$$

$$||x(t)-y||\le C\sum_{r=0}^{t}\alpha(r)+||x_0-y||,$$

where in the first relation we use the monotonicity of $\alpha(t)$ and the fact $|p(t)-t|\le T$, and in the second relation we use Eq. (13). By substituting the last two relations in the preceding inequality, we have

$$\sum_{l=1}^{m}\sum_{t\in M_{K,N}}\alpha(t)\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l(x(t))-f_l(y)\Big)\le C^2T\sum_{t=K}^{N}\alpha^2(t-T)$$

$$+\max\{C,G(y)\}\sum_{t=K}^{N}\Big(\alpha(t-T)-\alpha(t+T)\Big)\left(C\sum_{r=0}^{t}\alpha(r)+||x_0-y||\right). \quad (17)$$

Now we consider the second term on the right hand side of Eq. (16). For $t=K,\ldots,N$ we may have $j(t)\notin\{i(K),\ldots,i(N)\}$ possibly at times $t=K,\ldots,K+T-1$ and $t=N+1-T,\ldots,N$. Therefore, from the convexity of each f_l, the subgradient boundedness, and the fact $\sum_{l=1}^m\delta^l_{j(t)}=1$ for all t, we obtain

$$\sum_{l=1}^{m}\sum_{t\in P_{K,N}}\alpha(t)\delta^l_{j(t)}\Big(f_l(x(t))-f_l(y)\Big)\le C\sum_{t=K}^{K-1+T}\alpha(t)||x(t)-y||+C\sum_{t=N+1-T}^{N}\alpha(t)||x(t)-y||.$$

By using Eq. (13), the triangle inequality, and the monotonicity of $\alpha(t)$, we have

$$\begin{aligned} C\sum_{t=K}^{K-1+T}\alpha(t)||x(t)-y|| &\leq C\sum_{t=K}^{K-1+T}\alpha(t)\Big(||x(t)-x(K)||+||x(K)-y||\Big) \\ &\leq C^2T^2\alpha^2(K)+CT\alpha(K)||x(K)-y|| \\ &\leq C^2T^2\alpha^2(K)+\frac{\alpha(K)}{2}\Big(C^2T^2+||x(K)-y||^2\Big), \end{aligned}$$

where in the last inequality we exploit the fact $2ab \leq a^2+b^2$ for any scalars a and b. Similarly, it can be seen that

$$\begin{aligned} C\sum_{t=N+1-T}^{N}\alpha(t)||x(t)-y|| &\leq C\sum_{t=N+1-T}^{N}\alpha(t)\Big(||x(t)-x(N+1)||+||x(N+1)-y||\Big) \\ &\leq C^2T^2\alpha^2(N+1-T)+CT\alpha(N+1-T)||x(N+1)-y|| \\ &\leq C^2T^2\alpha^2(N+1-T) \\ &\quad +\frac{\alpha(N+1-T)}{2}\left(\beta C^2T^2+\frac{1}{\beta}||x(N+1)-y||^2\right), \end{aligned}$$

where the last inequality follows from the fact $2ab \leq \beta a^2+\frac{1}{\beta}b^2$ for any scalars a,b, and β with $\beta>0$. Therefore

$$\begin{aligned} \sum_{l=1}^{m}\sum_{t\in P_{K,N}}\alpha(t)\delta^l_{j(t)}\Big(f_l(x(t))-f_l(y)\Big) &\leq C^2T^2\Big(\alpha^2(K)+\alpha^2(N+1-T)\Big) \\ &\quad +\frac{C^2T^2}{2}\Big(\alpha(K)+\beta\alpha(N+1-T)\Big) \\ &\quad +\frac{1}{2}\left(\alpha(K)||x(K)-y||^2+\frac{1}{\beta}\alpha(N+1-T)||x(N+1)-y||^2\right). \quad (18) \end{aligned}$$

Finally, we estimate the last term in Eq. (16). For $t=K,\ldots,N$ we may have $i(t)\notin\{j(K),\ldots,j(N)\}$ possibly at times $t=K,\ldots,K+T-1$ and $t=N+1-T,\ldots,N$. Therefore, similar to the preceding analysis, it can be seen that

$$\begin{aligned} -\sum_{l=1}^{m}\sum_{t\in Q_{K,N}}\alpha(t)\delta^l_{i(t)}\Big(f_l(x(t))-f_l(y)\Big) &\leq G(y)CT^2\Big(\alpha^2(K)+\alpha^2(N+1-T)\Big) \\ &\quad +\frac{G^2(y)T^2}{2}\Big(\alpha(K)+\beta\alpha(N+1-T)\Big) \\ &\quad +\frac{1}{2}\left(\alpha(K)||x(K)-y||^2+\frac{1}{\beta}\alpha(N+1-T)||x(N+1)-y||^2\right), \quad (19) \end{aligned}$$

where $G(y)$ is given by Eq. (10). By substituting Eqs. (17)–(19) in the relation (16), and by using the definition of $c(y)$ [cf. Eq. (11)], we obtain Eq. (9). **Q.E.D.**

Lemma 4.2: Let Assumption 2.1 hold. Then, for the iterates $x(t)$ generated by the method, the following hold:

(a) For all $y \in X$, and k_0 and $\hat{k}$ with $\hat{k} > k_0$, we have

$$\left(1 - \frac{2}{\beta}\,\alpha(t_{\hat{k}} - W)\right)||x_{\hat{k}} - y||^2 \leq \Big(1 + 2\alpha(t_{k_0})\Big)||x_{k_0} - y||^2$$

$$-2\sum_{k=k_0}^{\hat{k}-1}\alpha(t_k)\Big(f(x_k) - f(y)\Big) + 2\tilde{C}\sum_{k=k_0}^{\hat{k}-1}\alpha^2(t_k - W)$$

$$+2K(y)\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\left(C\sum_{r=0}^{t_{k+1}}\alpha(r) + ||x_0 - y||\right)$$

$$+2c(y)\Big(\alpha^2(t_{k_0}) + \alpha(t_{k_0}) + \alpha^2(t_{\hat{k}} - W) + \beta\alpha(t_{\hat{k}} - W)\Big), \tag{20}$$

where $W = \max\{T, D\}$, β is an arbitrary positive scalar,

$$K(y) = mC + m\max\{C, G(y)\},$$

$$\tilde{C} = mC^2\left(\frac{1}{2} + m + 2D + T\right), \tag{21}$$

and $G(y)$ and $c(y)$ are defined by Eqs. (10) and (11), respectively.

(b) For for all $y \in X$ and $\hat{k} \geq 1$, we have

$$\frac{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)f(x_k)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)} \leq f(y) + \frac{\Big(1 + 2\alpha(0)\Big)||x_0 - y||^2}{2\sum_{k=0}^{\hat{k}-1}\alpha(t_k)} + \tilde{C}\frac{\sum_{k=0}^{\hat{k}-1}\alpha^2(t_k - W)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}$$

$$+K(y)\frac{\sum_{k=0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\Big(C\sum_{r=0}^{t_{k+1}}\alpha(r) + ||x_0 - y||\Big)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}$$

$$+c(y)\frac{\Big(\alpha^2(0) + \alpha(0) + \alpha^2(t_{\hat{k}} - W) + \beta\alpha(t_{\hat{k}} - W)\Big)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}, \tag{22}$$

where $\beta \geq 2\alpha(0)$.

Proof: (a) By using the convexity of each $f_{j(t)}$, the subgradient boundedness, the monotonicity of $\alpha(t)$, and the following relation [cf. Eq. (13)]

$$||x(t) - x(\hat{t})|| \leq C\sum_{s=\hat{t}}^{t-1}\alpha(s), \qquad \forall\ t, \hat{t},\ t \geq \hat{t},$$

we have for any $t \in \{t_k, \ldots, t_{k+1} - 1\}$

$$f_{j(t)}(x(t)) \geq f_{j(t)}(x_k) + g_{j(t)}(t_k)'\Big(x(t) - x_k\Big) \geq f_{j(t)}(x_k) - mC^2\alpha(t_k),$$

where $g_{j(t)}(t_k)$ is a subgradient of $f_{j(t)}$ at x_k. By substituting this relation in Eq. (8) [cf. Lemma 4.1] and by summing over $t = t_k, \ldots, t_{k+1} - 1$, we obtain

$$\begin{aligned} ||x_{k+1} - y||^2 \;\leq\; & ||x_k - y||^2 - 2\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ & + mC^2(1 + 2m + 4D)\alpha^2(t_k - D) \\ & + 2\sum_{l=1}^{m}\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big), \end{aligned} \tag{23}$$

where we also use $\alpha(t_k) \leq \alpha(t_k - D)$ and

$$\sum_{t=t_k}^{t_{k+1}-1} \alpha^2(t - D) \leq m\alpha^2(t_k - D), \qquad \sum_{t=t_k}^{t_{k+1}-1} \alpha(t) \leq m\alpha(t_k),$$

which follow from the monotonicity of $\alpha(t)$ and the fact $t_{k+1} - t_k = m$ for all k.

Now we estimate the second term on the right hand side in the inequality (23). For this we define

$$I_k^+(y) = \Big\{t \in \{t_k, \ldots, t_{k+1} - 1\} \mid f_{j(t)}(x_k) - f_{j(t)}(y) \geq 0\Big\},$$

$$I_k^-(y) = \{t_k, \ldots, t_{k+1} - 1\} \setminus I_k^+(y).$$

Since $\alpha(t_k) \geq \alpha(t) \geq \alpha(t_{k+1})$ for all t with $t_k \leq t < t_{k+1}$, we have for $t \in I_k^+(y)$

$$\alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \geq \alpha(t_{k+1})\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big),$$

and for $t \in I_k^-(y)$

$$\alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \geq \alpha(t_k)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big).$$

Hence for all t with $t_k \leq t < t_{k+1}$

$$\begin{aligned} \sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \;\geq\; & \alpha(t_{k+1}) \sum_{t\in I_k^+(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ & + \alpha(t_k) \sum_{t\in I_k^-(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ = \;& \alpha(t_k) \sum_{t=t_k}^{t_{k+1}-1} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ - \;& \Big(\alpha(t_k) - \alpha(t_{k+1})\Big) \sum_{t\in I_k^+(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big). \end{aligned} \tag{24}$$

Furthermore, by using the convexity of each $f_{j(t)}$, the subgradient boundedness, and Eq. (13), we can see that

$$f_{j(t)}(x_k) - f_{j(t)}(y) \leq C\Big(||x_k - x_0|| + ||x_0 - y||\Big) \leq C\left(C\sum_{r=0}^{t_k}\alpha(r) + ||x_0 - y||\right),$$

and, since the cardinality of $I_k^+(y)$ is at most m, we obtain

$$\sum_{t\in I_k^+(y)}\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \leq mC\left(C\sum_{r=0}^{t_k}\alpha(r) + ||x_0 - y||\right).$$

For the cyclic rule we have $\{j(t_k), \ldots, j(t_{k+1}-1)\} = \{1, \ldots, m\}$, so that

$$\sum_{t=t_k}^{t_{k+1}-1}\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) = f(x_k) - f(y).$$

By using the last two relations in Eq. (24), we obtain

$$\begin{aligned}\sum_{t=t_k}^{t_{k+1}-1}\alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \;&\geq\; \alpha(t_k)\Big(f(x_k) - f(y)\Big)\\ &\quad -mC\Big(\alpha(t_k) - \alpha(t_{k+1})\Big)\left(C\sum_{r=0}^{t_k}\alpha(r) + ||x_0 - y||\right),\end{aligned}$$

which when substituted in Eq. (23) yields for all $y \in X$ and k

$$\begin{aligned}||x_{k+1} - y||^2 \;&\leq\; ||x_k - y||^2 - 2\alpha(t_k)\Big(f(x_k) - f(y)\Big) + mC^2(1 + 2m + 4D)\alpha^2(t_k - D)\\ &\quad +2mC\Big(\alpha(t_k) - \alpha(t_{k+1})\Big)\left(C\sum_{r=0}^{t_k}\alpha(r) + ||x_0 - y||\right)\\ &\quad +2\sum_{l=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big).\end{aligned}$$

By summing these inequalities over $k = k_0, \ldots, \hat{k} - 1$, and by using the facts $t_k < t_{k+1}$, $\alpha(t_k) \leq \alpha(t_k - W)$, $\alpha(t_{k+1}) \geq \alpha(t_{k+1} + W)$, $\alpha^2(t_k - D) \leq \alpha^2(t_k - W)$, we obtain

$$\begin{aligned}||x_{\hat{k}} - y||^2 \;&\leq\; ||x_{k_0} - y||^2 - 2\sum_{k=k_0}^{\hat{k}-1}\alpha(t_k)\Big(f(x_k) - f(y)\Big)\\ &\quad +mC^2(1 + 2m + 4D)\sum_{k=k_0}^{\hat{k}-1}\alpha^2(t_k - W)\\ &\quad +2mC\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\left(C\sum_{r=0}^{t_{k+1}}\alpha(r) + ||x_0 - y||\right)\\ &\quad +2\sum_{k=k_0}^{\hat{k}-1}\sum_{l=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big),\end{aligned} \tag{25}$$

with $W = \max\{D, T\}$. Note that the last term in the preceding relation can be written as the sum over $l = 1, \ldots, m$ and over $t = t_{k_0}, \ldots, t_{\hat{k}} - 1$, so that by using Eq. (9) of Lemma 4.1 where $K = t_{k_0}$ and $N = t_{\hat{k}} - 1$, and the monotonicity of $\alpha(t)$, we have

$$\sum_{l=1}^{m} \sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big) \leq C^2 T \sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha^2(t - W)$$

$$+ \max\{C, G(y)\} \sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \Big(\alpha(t - W) - \alpha(t + W)\Big) \left(C \sum_{r=0}^{t} \alpha(r) + ||x_0 - y||\right)$$

$$+ c(y)\Big(\alpha^2(t_{k_0}) + \alpha(t_{k_0}) + \alpha^2(t_{\hat{k}} - W) + \beta\alpha(t_{\hat{k}} - W)\Big)$$

$$+ \left(\alpha(t_{k_0})||x(t_{k_0}) - y||^2 + \frac{1}{\beta}\alpha(t_{\hat{k}} - W)||x_{\hat{k}} - y||^2\right). \tag{26}$$

Furthermore, by the monotonicity of $\alpha(t)$, it follows that

$$\sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha^2(t - W) = \sum_{k=k_0}^{\hat{k}-1} \sum_{t=t_k}^{t_{k+1}-1} \alpha^2(t - W) \leq m \sum_{k=k_0}^{\hat{k}-1} \alpha^2(t_k - W),$$

$$\sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \Big(\alpha(t - W) - \alpha(t + W)\Big) \left(C \sum_{r=0}^{t} \alpha(r) + ||x_0 - y||\right) \leq$$

$$m \sum_{k=k_0}^{\hat{k}-1} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \left(C \sum_{r=0}^{t_{k+1}} \alpha(r) + ||x_0 - y||\right).$$

The desired relation (20) follows by substituting the last two inequalities in Eq. (26) and by using the resulting relation in Eq. (25).

(b) The desired relation (22) follows from Eq. (20), by setting $k_0 = 0$, by dividing with $2\sum_{k=0}^{\hat{k}-1} \alpha(t_k)$, and by using the fact $\beta \geq 2\alpha(0) \geq 2\alpha(t)$ for all t. **Q.E.D.**

Now we prove Prop. 2.1.

Proof of Prop. 2.1: We prove (a) and (b) simultaneously. Since $\alpha(t) = \alpha$ for $t \in (-\infty, \infty)$ [recall that we defined $\alpha(t) = \alpha(0)$ for $t < 0$], from Lemma 4.2(b), we obtain for all $y \in X$ and $\hat{k} \geq 1$

$$\frac{1}{\hat{k}} \sum_{k=0}^{\hat{k}-1} f(x_k) \leq f(y) + \frac{(1 + 2\alpha)||x_0 - y||^2}{2\alpha\hat{k}} + \tilde{C}\alpha + c(y)\frac{1 + 2\alpha + \beta}{\alpha\hat{k}}.$$

By letting $\hat{k} \to \infty$ and by using the following inequality

$$\liminf_{\hat{k}\to\infty} f(x_{\hat{k}}) \leq \liminf_{\hat{k}\to\infty} \frac{1}{\hat{k}} \sum_{k=0}^{\hat{k}-1} f(x_k),$$

we have for all $y \in X$

$$\liminf_{\hat{k}\to\infty} f(x_{\hat{k}}) \leq f(y) + \tilde{C}\alpha,$$

from which the results (a) and (b) follow by taking the minimum over $y \in X$. **Q.E.D.**

In the proofs of Props. 2.2 and 2.3 we use properties of the stepsize given in the following lemma.

Lemma 4.3: Let the stepsize $\alpha(t)$ satisfy Assumption 2.2. Then we have

$$\lim_{k\to\infty} \frac{\alpha^2(t_k - W)}{\alpha(t_k)} = 0, \qquad \lim_{k\to\infty} \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{t=0}^{t_{k+1}} \alpha(t) = 0,$$

$$\sum_{k=0}^{\infty} \alpha(t_k) = \infty, \qquad \sum_{k=0}^{\infty} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) < \infty,$$

where $t_k = mk$ and W is a nonnegative integer. In addition, for $\frac{1}{2} < q \leq 1$ we have

$$\sum_{k=0}^{\infty} \alpha^2(t_k - W) < \infty, \qquad \sum_{k=0}^{\infty} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \sum_{t=0}^{t_{k+1}} \alpha(t) < \infty.$$

Proof: Let $0 < q \leq 1$. The stepsize $\alpha(t)$ is smallest when $S = 1$, so that

$$\sum_{k=0}^{\infty} \alpha(t_k) \geq r_0 \sum_{k=0}^{\infty} \frac{1}{(km + r_1)^q} = \infty.$$

Let $\{l_k\}$ be a sequence of nonnegative integers such that for all k

$$\alpha(t_k - W) = \frac{r_0}{(l_k + r_1)^q}. \tag{27}$$

Note that $l_k \to \infty$ as $k \to \infty$. Given the value of $\alpha(t_k - W)$, the values of $\alpha(t_k)$ and $\alpha(t_{k+1} + W)$ are smallest if we decrease the stepsize $\alpha(t)$ at each time t for $t > t_k - W$. Therefore

$$\alpha(t_k) \geq \frac{r_0}{(l_k + W + r_1)^q}, \tag{28}$$

$$\alpha(t_{k+1} + W) \geq \frac{r_0}{(l_k + m + 2W + r_1)^q}, \tag{29}$$

where in the last inequality above we use the fact $t_k = mk$. By combining Eqs. (27) and (28), we see that

$$\lim_{k\to\infty} \frac{\alpha^2(t_k - W)}{\alpha(t_k)} = 0.$$

Next from Eqs. (27) and (29) we obtain

$$\begin{aligned} \alpha(t_k - W) - \alpha(t_{k+1} + W) &= r_0 \frac{(l_k + m + 2W + r_1)^q - (l_k + r_1)^q}{(l_k + r_1)^q (l_k + m + 2W + r_1)^q} \\ &\le \frac{r_0 q (m + 2W)}{(l_k + r_1)(l_k + W + r_1)^q}, \end{aligned} \tag{30}$$

where in the last inequality above we exploit the facts $l_k + m + 2W + r_1 \ge l_k + W + r_1$ and

$$b^q - a^q = q \int_a^b \frac{dx}{x^{1-q}} \le \frac{q}{a^{1-q}} \int_a^b dx = \frac{q(b-a)}{a^{1-q}},$$

for all b and a with $b \ge a > 0$, and $0 < q \le 1$. In particular, the relation (30) implies that

$$\alpha(t_k - W) - \alpha(t_{k+1} + W) \le \frac{r_0 q (m + 2W)}{(l_k + r_1)^{1+q}}, \tag{31}$$

so that $\sum_{k=0}^{\infty} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) < \infty$. Furthermore, by combining Eqs. (28) and (30), we obtain for all k

$$\frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \le \frac{q(m + 2W)}{l_k + r_1}. \tag{32}$$

Now we estimate $\sum_{t=0}^{t_{k+1}} \alpha(t)$. By using the definition and the monotonicity of $\alpha(t)$, and Eq. (27) we have for all k large enough (so that $t_k - W > 0$)

$$\sum_{t=0}^{t_{k+1}} \alpha(t) \le \sum_{t=0}^{t_k - W} \alpha(t) + (1 + m + W)\alpha(t_k - W) \le \sum_{l=0}^{l_k} \frac{S r_0}{(l + r_1)^q} + (1 + m + W)\alpha(t_k - W).$$

Since

$$\sum_{l=0}^{l_k} \frac{1}{(l + r_1)^q} \le \begin{cases} \frac{1}{r_1} + \ln(l_k + r_1) & \text{if } q = 1, \\ \frac{1}{r_1^q} + \frac{(l_k + r_1)^{1-q}}{1-q} & \text{if } 0 < q < 1, \end{cases} \tag{33}$$

from the preceding relation we obtain

$$\sum_{t=0}^{t_{k+1}} \alpha(t) \le u_k, \tag{34}$$

where

$$u_k = \begin{cases} O\Big(\ln(l_k + r_1)\Big) & \text{if } q = 1, \\ O\Big((l_k + r_1)^{1-q}\Big) & \text{if } 0 < q < 1. \end{cases} \tag{35}$$

This together with Eq. (32) implies that

$$\lim_{k\to\infty} \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{t=0}^{t_{k+1}} \alpha(t) = 0.$$

Now, let $\frac{1}{2} < q \leq 1$. Then, by using the definition of $\alpha(t)$, we have for K large enough (so that $t_k - W > 0$)

$$\sum_{k=K}^{\infty} \alpha^2(t_k - W) \leq \sum_{t=K}^{\infty} \alpha^2(t - W) \leq \sum_{s=0}^{\infty} \alpha^2(s) \leq \sum_{l=0}^{\infty} \frac{S r_1^2}{(l + r_1)^{2q}},$$

implying that $\sum_{k=0}^{\infty} \alpha^2(t_k - W)$ is finite. Furthermore, by combining Eqs. (31), (34), and (35), we obtain

$$\alpha(t_k - W) - \alpha(t_{k+1} + W) \sum_{t=0}^{t_{k+1}} \alpha(t) \leq v_k,$$

where

$$v_k = \begin{cases} O\left(\frac{\ln(l_k + r_1)}{(l_k + r_1)^2}\right) & \text{if } q = 1, \\ O\left(\frac{1}{(l_k + r_1)^{2q}}\right) & \text{if } 0 < q < 1. \end{cases}$$

Hence $\sum_{k=0}^{\infty} \alpha(t_k - W) - \alpha(t_{k+1} + W) \sum_{t=0}^{t_{k+1}} \alpha(t)$ is finite. **Q.E.D.**

We are now ready to prove Props. 2.2 and 2.3.

Proof of Prop. 2.2: It suffices to show that $\liminf_{k\to\infty} f(x_k) = f^*$. For this we need the following two relations

$$\liminf_{N\to\infty} b_N \leq \liminf_{N\to\infty} \frac{\sum_{k=0}^{N-1} c_k b_k}{\sum_{k=0}^{N-1} c_k}, \tag{36}$$

$$\limsup_{N\to\infty} \frac{\sum_{k=0}^{N-1} c_k b_k}{\sum_{k=0}^{N-1} c_k} \leq \limsup_{N\to\infty} b_N, \tag{37}$$

which hold for any two scalar sequences $\{b_k\}$ and $\{c_k\}$ with $c_k > 0$ for all k and $\sum_{k=0}^{\infty} c_k = \infty$ (see Lemma 2.1 of Kiwiel [14]). By letting $\hat{k} \to \infty$ in Eq. (22) of Lemma 4.2, by using Lemma 4.3 and the relation (37), where $c_k = \alpha(t_k)$ and

$$b_k = \frac{\tilde{C}\alpha^2(t_k - W)}{\alpha(t_k)} + K(y) \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{t=0}^{t_{k+1}} \alpha(t),$$

from Eq. (22) it can be seen that for all $y \in X$

$$\liminf_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha(t_k) f(x_k)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} \leq f(y).$$

By using the relation (36) with $c_k = \alpha(t_k)$ and $b_k = f(x_k)$, from the preceding relation we obtain for all $y \in X$

$$\liminf_{\hat{k}\to\infty} f(x_{\hat{k}}) \leq f(y),$$

from which the result follows by taking the minimum over $y \in X$. **Q.E.D.**

Proof of Prop. 2.3: By letting $\beta = 1$ and $y = x^*$ for some $x^* \in X^*$, and by dropping the nonpositive term involving $f(x_k) - f^*$, from Eq. (20) of Lemma 4.2, we obtain for all $x^* \in X^*$ and $\hat{k} > k_0$

$$\Big(1 - 2\alpha(t_{\hat{k}} - W)\Big)||x_{\hat{k}} - x^*||^2 \leq \Big(1 + 2\alpha(t_{k_0})\Big)||x_{k_0} - x^*||^2 + 2\tilde{C} \sum_{k=k_0}^{\hat{k}-1} \alpha^2(t_k - W)$$

$$+2K(x^*) \sum_{k=k_0}^{\hat{k}-1} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \left(C \sum_{r=0}^{t_{k+1}} \alpha(r) + ||x_0 - x^*||\right)$$

$$+2c(x^*)\Big(\alpha^2(t_{k_0}) + \alpha(t_{k_0}) + \alpha^2(t_{\hat{k}} - W) + \alpha(t_{\hat{k}} - W)\Big), \tag{38}$$

As $\hat{k} \to \infty$, by applying Lemma 4.3, from the preceding relation it follows that

$$\limsup_{\hat{k}\to\infty} ||x_{\hat{k}} - x^*|| < \infty,$$

i.e., $\{x_k\}$ is bounded. Furthermore, according to Prop. 2.2, we have

$$\liminf_{k\to\infty} f(x_k) = f^*,$$

which, by the continuity of f and the boundedness of $\{x_k\}$, implies the existence of a subsequence $\{x_{k_j}\} \subset \{x_k\}$ and a point $\hat{x} \in X^*$ such that

$$\lim_{j\to\infty} ||x_{k_j} - \hat{x}|| = 0.$$

Next, in Eq. (38) we let $x^* = \hat{x}$ and $k_0 = k_j$ for some j. Then by first letting $\hat{k} \to \infty$ and then $j \to \infty$, and by using Lemma 4.3 and the fact $x_{k_j} \to \hat{x}$, we obtain

$$\limsup_{\hat{k}\to\infty} ||x_{\hat{k}} - \hat{x}|| = 0.$$

Q.E.D.

5. Convergence Proofs for Random Selection Rule

In this section we give proofs of Props. 3.1 and 3.2. The proofs rely on the martingale convergence theorem, as stated for example in Gallager [9], p. 256.

Theorem 5.1: (*Martingale Convergence Theorem*) Let $\{Z_k : k = 0, 1, \ldots\}$ be a martingale and assume that there is some positive scalar M such that $E\left\{Z_k^2\right\} \leq M$ for all k. Then there is a random variable Z such that, for all sample sequences except a set of probability 0, we have

$$\lim_{k\to\infty} Z_k = Z.$$

In the proofs we also use some properties of $\alpha(t)$ and $x(t)$ that are given, respectively, in Lemmas 5.1 and 5.2 below.

Lemma 5.1: Let Assumption 2.2 hold with $\frac{3}{4} < q \leq 1$. Then we have

$$\sum_{t=0}^{\infty} \alpha^2(t-W) < \infty, \qquad \sum_{t=0}^{\infty} \Delta(t) < \infty, \qquad \sum_{t=0}^{\infty} \alpha(t) = \infty,$$

$$\sum_{t=0}^{\infty} \Delta(t) \sum_{r=0}^{t} \alpha(r) < \infty, \qquad \sum_{t=0}^{\infty} \alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2 < \infty, \tag{39}$$

where $\Delta(t) = \alpha(t-T) - \alpha(t+T)$ and W is a nonnegative integer.

Proof: We show the last relation in Eq. (39). The rest can be shown similar to the proof of Lemma 4.3. Note that $\alpha(t)$ is largest when we change the step every S iterations, i.e., $\sigma_{l+1} - \sigma_l = S$ for all l, so that

$$\alpha(t) \leq \frac{r_0}{(l+r_1)^q}, \qquad t \in \{lS, \ldots, (l+1)S-1\}, \quad l = 0, 1, \ldots,$$

and consequently [cf. Eq. (33)]

$$\sum_{r=0}^{t} \alpha(r) \leq Sr_0 \sum_{k=0}^{l} \frac{1}{(k+r_1)^q} \leq \begin{cases} Sr_0 \left(\frac{1}{r_1} + \ln(l+r_1)\right) & \text{if } q = 1, \\ Sr_0 \left(\frac{1}{r_1^q} + \frac{(l+r_1)^{1-q}}{1-q}\right) & \text{if } q < 1. \end{cases}$$

Therefore $\alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2 \leq w_l$ for $t \in \{lS, \ldots, (l+1)S-1\}$, where w_l is of the order $\frac{\ln^2(l+1+r_1)}{(l+r_1)^2}$ for $q = 1$ and of the order $\frac{1}{(l+r_1)^{4q-2}}$ for $q < 1$. Hence $\sum_{t=0}^{\infty} \alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2$ is finite for $q > \frac{3}{4}$. **Q.E.D.**

Lemma 5.2: Let Assumption 3.1 hold.

(a) For all $y \in X$ and t, we have

$$\begin{aligned} ||x(t+1)-y||^2 \;\leq\; & ||x(t)-y||^2 - \frac{2\alpha(t)}{m}\Big(f(x(t))-f(y)\Big) + 2\Big(z_y(t)-z_y(t-1)\Big) \\ & +C^2(1+4D)\alpha^2(t-D) \\ & +2\alpha(t)\sum_{l=1}^{m}\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l(x(t))-f_l(y)\Big), \end{aligned} \tag{40}$$

where δ^l_i is the Kronecker symbol, $z_y(-1)=0$, and for $t \geq 0$

$$z_y(t) = \sum_{r=0}^{t}\alpha(r)\left(\frac{1}{m}\Big(f(x(r))-f(y)\Big) - \Big(f_{j(r)}(x(r))-f_{j(r)}(y)\Big)\right). \tag{41}$$

(b) For all $y \in X$, and N and K with $N \geq K$, we have

$$\begin{aligned} ||x(N+1)-y||^2 \;\leq\; & ||x(K)-y||^2 - \frac{2}{m}\sum_{t=K}^{N}\alpha(t)\Big(f(x(t))-f(y)\Big) \\ & +2\Big(z_y(N)-z_y(K-1)\Big) \\ & +C^2\Big(1+4D+2T\Big)\sum_{t=K}^{N}\alpha^2(t-W) \\ & +2\max\{G(y),C\}\sum_{t=K}^{N}\Delta(t)\left(C\sum_{r=0}^{t}\alpha(r)+||x(0)-y||\right) \\ & +2c(y)\Big(\alpha^2(K)+\alpha(K)+\alpha^2(N+1-T)+\alpha(N+1-T)\Big) \\ & +2\Big(\alpha(K)||x(K)-y||^2+\alpha(N+1-T)||x(N+1)-y||^2\Big), \end{aligned}$$

where $W=\max\{D,T\}$, $G(y)$ and $c(y)$ are given by Eqs. (10) and (11), respectively, and $\Delta(t)=\alpha(t-T)-\alpha(t+T)$ for all t.

(c) For all $y \in X$, the sequence $\{z_y(t)\}$ defined by Eq. (41) is a convergent martingale.

Proof: (a) The relation (40) follows from Eq. (8) of Lemma 4.1 by adding and subtracting $\frac{2\alpha(t)}{m}\Big(f(x_k)-f(y)\Big)$, and by using the definition of $z_y(t)$ [cf. Eq. (41)].

(b) Summing the inequalities (40) over $t=K,\ldots,N$ yields for all $y \in X$

$$\begin{aligned} ||x(N+1)-y||^2 \;\leq\; & ||x(K)-y||^2 - \frac{2}{m}\sum_{t=K}^{N}\alpha(t)\Big(f(x(t))-f(y)\Big) \\ & +2\Big(z_y(N)-z_y(K-1)\Big) + C^2(1+4D)\sum_{t=K}^{N}\alpha^2(t-D) \\ & +2\sum_{t=K}^{N}\sum_{l=1}^{m}\alpha(t)\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l(x(t))-f_l(y)\Big). \end{aligned}$$

The desired relation follows from the preceding relation, by using Eq. (9) of Lemma 4.1 with $\beta=1$ and $x_0=x(0)$.

(c) Let $y \in X$ be fixed. First, we show that the sequence $\{z_y(t)\}$ is a martingale. By using the definition of $z_y(t)$ [cf. Eq. (41)], we have

$$\begin{aligned} E\Big\{z_y(t) \mid z_y(t-1)\Big\} &= z_y(t-1) \\ &\quad +\alpha(t)E\left\{\frac{1}{m}\Big(f(x(t))-f(y)\Big)-\Big(f_{j(t)}(x(t))-f_{j(t)}(y)\Big)\right\} \\ &= z_y(t-1), \end{aligned}$$

where in the last equality we use the iterated expectation rule and

$$E\left\{\frac{1}{m}\Big(f(x(t))-f(y)\Big)-\Big(f_{j(t)}(x(t))-f_{j(t)}(y)\Big)\,\Big|\, x(t)\right\}=0,$$

which follows from the properties of $\{j(t)\}$ [cf. Assumption 3.1(c)]. Hence $z_y(t)$ is indeed a martingale.

To apply the martingale convergence theorem, which guarantees the convergence of $z_y(t)$, we have to prove that $E\Big\{z_y^2(t)\Big\}$ is bounded. From the definition of $z_y(t)$ [cf. Eq. (41)] it follows that

$$E\Big\{z_y^2(t)\Big\}=\sum_{r=0}^{t}\alpha^2(r)E\left\{\left(\frac{1}{m}\Big(f(x(r))-f(y)\Big)-\Big(f_{j(r)}(x(r))-f_{j(r)}(y)\Big)\right)^2\right\}. \tag{42}$$

This is because the expected values of the cross terms appearing in $z_y^2(t)$ are equal to 0, which can be seen by using the iterated expectation rule [i.e., by conditioning on the values $x(s)$ and $x(r)$ for $s, r \le t$] and by exploiting the properties of $j(r)$ [cf. Assumption 3.1(c)]. Furthermore, by using convexity of each f_i, the triangle inequality, and the following relation [cf. Eq. (13)]

$$||x(t)-x(\hat{t})|| \le C\sum_{s=\hat{t}}^{t-1}\alpha(s), \qquad \forall\ t,\hat{t},\ t\ge\hat{t},$$

for every r we have

$$\begin{aligned} &\left(\frac{1}{m}\Big(f(x(r))-f(y)\Big)-\Big(f_{j(r)}(x(r))-f_{j(r)}(y)\Big)\right)^2 \\ &\qquad \le \frac{2}{m^2}\Big(f(x(r))-f(y)\Big)^2+2\Big(f_{j(r)}(x(r))-f_{j(r)}(y)\Big)^2 \\ &\qquad \le 4\Big(\max\{G(y),C\}\Big)^2||x(r)-y||^2 \\ &\qquad \le 4\Big(\max\{G(y),C\}\Big)^2\left(C\sum_{s=0}^{r}\alpha(s)+||x(0)-y||\right)^2 \\ &\qquad \le 8\Big(\max\{G(y),C\}\Big)^2\left[C^2\left(\sum_{s=0}^{r}\alpha(s)\right)^2+||x(0)-y||^2\right]. \end{aligned}$$

By using this inequality and Lemma 5.1 in Eq. (42), we see that $E\Big\{z_y^2(t)\Big\}$ is bounded. Thus, according to the martingale convergence theorem, as $t \to \infty$, the sequence $\{z_y(t)\}$ converges to some random variable with probability 1. **Q.E.D.**

Next we prove Prop. 3.1.

Proof of Prop. 3.1: First we consider the case where f^* is finite. Let $\epsilon > 0$ be arbitrary and let $\hat{y} \in X$ be such that

$$f(\hat{y}) \leq f^* + \epsilon.$$

Fix a sample path, denoted by $\mathcal{P}$, for which the martingale $\{z_{\hat{y}}(t)\}$ is convergent [cf. Lemma 5.2(c)]. From Lemma 5.2(b), where $K = 0$ and $y = \hat{y}$, we have for the path $\mathcal{P}$ and sufficiently large N

$$\begin{aligned}\frac{2}{m}\sum_{t=0}^{N}\alpha(t)\Big(f(x(t)) - f(\hat{y})\Big) \quad \leq \quad & \Big(1 + 2\alpha(0)\Big)||x(0) - \hat{y}||^2 + 2z_{\hat{y}}(N) \\ & + C^2\Big(1 + 4D + 2T\Big)\sum_{t=0}^{N}\alpha^2(t - W) \\ & + 2\max\{C, G(\hat{y})\}\sum_{t=0}^{N}\Delta(t)\left(C\sum_{r=0}^{t}\alpha(r) + ||x(0) - \hat{y}||\right) \\ & + 2c(\hat{y})\Big(\alpha^2(0) + \alpha(0) + \alpha^2(N + 1 - T) + \alpha(N + 1 - T)\Big),\end{aligned}$$

where we use the fact $z_{\hat{y}}(-1) = 0$ and we take $N \geq k_0$ with k_0 such that $1 - 2m\alpha(k - T) \geq 0$ for all $k \geq k_0$. Since $z_{\hat{y}}(N)$ converges, by dividing the above inequality with $\frac{2}{m}\sum_{t=0}^{N}\alpha(t)$, by letting $N \to \infty$, and by using Lemma 5.1, we obtain

$$\liminf_{N\to\infty}\frac{\sum_{t=0}^{N}\alpha(t)f(x(t))}{\sum_{t=0}^{N}\alpha(t)} \leq f(\hat{y}).$$

By using the facts $f(\hat{y}) \leq f^* + \epsilon$ and

$$\liminf_{N\to\infty} f(x(N+1)) \leq \liminf_{N\to\infty}\frac{\sum_{t=0}^{N}\alpha(t)f(x(t))}{\sum_{t=0}^{N}\alpha(t)}$$

[cf. Eq. (36)], we have for the path $\mathcal{P}$

$$\liminf_{N\to\infty} f(x(N+1)) \leq f^* + \epsilon.$$

Thus $\liminf_{t\to\infty} f(x(t)) \leq f^* + \epsilon$ with probability 1, and since ϵ is arbitrary, it follows that

$$\liminf_{t\to\infty} f(x(t)) = f^*.$$

In the case where $f^* = -\infty$, we choose an arbitrarily large positive integer M and a point $\hat{y} \in X$ such that $f(\hat{y}) \leq -M$, and use the same line of analysis as in the preceding case. **Q.E.D.**

Now we present a proof of Prop. 3.2

Proof of Prop. 3.2: Let d be the dimension of X^*. Then there exist $d+1$ distinct points $y_0, \ldots, y_d \in X^*$ that are in general position, i.e., they are such that the vectors $y_1 - y_0, \ldots, y_d - y_0$ are linearly independent. According to Lemma 5.2(c), for each $s = 0, \ldots, d$ the martingale $z_s(t) = z_{y_s}(t)$ converges, as $t \to \infty$, to some random variable with probability 1.

Now, fix a sample path, denoted by $\mathcal{P}$, for which

$$\liminf_{t\to\infty} f(x(t)) = f^* \tag{43}$$

(cf. Prop. 3.1) and every martingale $z_s(t)$ is convergent. Furthermore, fix an $s \in \{0, \ldots, d\}$. Let K_0 be a positive integer large enough so that

$$1 - 2\alpha(K - T) > 0, \qquad \forall\ K \geq K_0.$$

By using Lemma 5.2(b) with $y = y_s$ and $N > K \geq K_0$, and by dropping the nonnegative term involving $f(x(t)) - f(y_s)$, we obtain

$$\begin{aligned}\Big(1 - 2\alpha(N+1-T)\Big)||x(N+1) - y_s||^2 \leq \Big(1 + 2\alpha(K)\Big)||x(K) - y_s||^2 \\ +2\Big(z_s(N) - z_s(K-1)\Big) + C^2\Big(1 + 4D + 2T\Big)\sum_{t=K}^{N}\alpha^2(t-W) \\ +2\max\{G(y_s), C\}\sum_{t=K}^{N}\Delta(t)\left(C\sum_{r=0}^{t}\alpha(r) + ||x(0) - y_s||\right) \\ +2c(y_s)\Big(\alpha^2(K) + \alpha(K) + \alpha^2(N+1-T) + \alpha(N+1-T)\Big).\end{aligned}$$

By using Lemma 5.1 and the convergence of $\{z_s(t)\}$, from the preceding relation we obtain

$$\limsup_{N\to\infty} ||x(N+1) - y_s||^2 \leq \liminf_{K\to\infty} ||x(K) - y_s||^2.$$

Hence $\lim_{t\to\infty} ||x(t) - y_s||^2$ exists for the path $\mathcal{P}$ and all $s = 0, \ldots, d$.

Let us define for $s = 0, \ldots, d$

$$\beta_s = \lim_{t\to\infty} ||x(t) - y_s||. \tag{44}$$

Because $y_0, \ldots, y_d$ are distinct points, at most one β_s can be zero, in which case for the path $\mathcal{P}$ we have $\lim_{t\to\infty} ||x(t) - y_s|| = 0$, and we are done. Now, let all the scalars β_s be positive. Without loss of generality, we may assume that all β_s are equal, for otherwise we can use a linear transformation of the space $\Re^n$ such that in the transformed space the values $\hat{\beta}_0, \ldots, \hat{\beta}_d$ corresponding to $\beta_0, \ldots, \beta_d$, respectively, are all equal. Therefore we set $\beta_s = \beta$ for all s, so that every limit point of $\{x(t)\}$ is at the same distance β from

each of the points $y_0, \ldots, y_d$. Let $\mathcal{M}_d$ be the d-dimensional linear manifold containing X^*. Because $y_0, \ldots, y_d$ are in general position, the set of all points that are at the same distance from each y_s is an $n-d$-dimensional linear manifold $\mathcal{M}_{n-d}$. Hence all limit points of $\{x(t)\}$ lie in the manifold $\mathcal{M}_{n-d}$.

In view of Eqs. (43) and (44), $\{x(t)\}$ must have a limit point $\tilde{x}$ for which $f(\tilde{x}) = f^*$. Therefore $\tilde{x} \in X^*$ implying that $\tilde{x} \in \mathcal{M}_d \cap \mathcal{M}_{n-d}$, and since $\mathcal{M}_d$ is orthogonal to $\mathcal{M}_{n-d}$, $\tilde{x}$ is the unique limit point of $\{x(t)\}$ for the path $\mathcal{P}$. Hence $\{x(t)\}$ converges to some optimal solution with probability 1. **Q.E.D.**

Remark 5.1: If the set X is compact, then it can be shown that

$$E\Big\{z_y^2(t)\Big\} \leq 4\Big(\max\{G(y), C\}\Big)^2 \sup_{x \in X} ||x-y||^2 \sum_{r=0}^{t} \alpha^2(r).$$

In this case, the martingale $\{z_y(t)\}$ is convergent for $\frac{1}{2} < q \leq 1$ in Assumption 3.1(b), so that the result of Prop. 3.2 holds under Assumption 3.1(b) with $\frac{1}{2} < q \leq 1$ instead of $\frac{3}{4} < q \leq 1$.

REFERENCES

1. A. Ben-Tal, T. Margalit, and A. Nemirovski, The Ordered Subsets Mirror Descent Optimization Method and its Use for the Positron Emission Tomography Reconstruction, submitted to the Proceedings of the March 2000 Haifa Workshop "Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications", D. Butnariu, Y. Censor, and S. Reich, Eds., *Studies in Computational Mathematics*, Elsevier, Amsterdam.
2. A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations* (Springer-Verlag, New York, 1990).
3. D. P. Bertsekas, A New Class of Incremental Gradient Methods for Least Squares Problems, *SIAM J. on Optimization* **7** (1997) 913–926.
4. D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods* (Prentice-Hall, Inc., 1989).
5. D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming* (Athena Scientific, Belmont, Massachusetts, 1996).
6. D. P. Bertsekas and J. N. Tsitsiklis, Gradient Convergence in Gradient Methods, *SIAM J. on Optimization* **3** (2000) 627–642.
7. V. S. Borkar, Asynchronous Stochastic Approximation, *SIAM J. on Optimization* **36** (1998) 840–851.
8. A. A. Gaivoronski, Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks, *Opt. Meth. and Software* **4** (1994) 117–134.
9. R. G. Gallager, *Discrete Stochastic Processes* (Kluwer Academic Publishers, 1996).
10. L. Grippo, A Class of Unconstrained Minimization Methods for Neural Network Training, *Opt. Meth. and Software* **4** (1994) 135–150.
11. C. A. Kaskavelis and M. C. Caramanis, Efficient Lagrangian Relaxation Algorithms for

Industry Size Job-Shop Scheduling Problems, *IIE Trans. on Scheduling and Logistics* **30** (1998) 1085–1097.

12. V. M. Kibardin, Decomposition into Functions in the Minimization Problem, *Automation and Remote Control* **40** (1980) 1311–1323.
13. K. C. Kiwiel and P. O. Lindberg, Parallel Subgradient Methods for Convex Optimization, submitted to the Proceedings of the March 2000 Haifa Workshop "Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications", D. Butnariu, Y. Censor, and S. Reich, Eds., *Studies in Computational Mathematics*, Elsevier, Amsterdam.
14. K. C. Kiwiel, Convergence of Approximate and Incremental Subgradient Methods for Convex Optimization, submitted to *SIAM J. on Optimization*.
15. H. J. Kushner and G. Yin, it Stochastic Approximation Algorithms and Applications (Springer-Verlag, New York, 1997).
16. Z. Q. Luo, On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks, *Neural Computation* **3** (1991) 226–245.
17. Z. Q. Luo and P. Tseng, Analysis of an Approximate Gradient Projection Method with Applications to the Backpropagation Algorithm, *Opt. Meth. and Software* **4** (1994) 85–101.
18. O. L. Mangasarian and M. V. Solodov, Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization, *Opt. Meth. and Software* **4** (1994) 103–116.
19. A. Nedić and D. P. Bertsekas, Incremental Subgradient Methods for Nondifferentiable Optimization, *Lab. for Info. and Decision Systems Report LIDS-P-2460* (Massachusetts Institute of Technology, Cambridge, MA, 1999).
20. A. Nedić and D. P. Bertsekas, Incremental Subgradient Methods for Nondifferentiable Optimization, submitted to *SIAM J. Optimization*.
21. A. Nedić and D. P. Bertsekas, Convergence Rate of Incremental Subgradient Algorithms, *Lab. for Info. and Decision Systems Report LIDS-P-2475* (Massachusetts Institute of Technology, Cambridge, MA, 2000), to appear in *Stochastic Optimization: Algorithms and Applications*, Eds. S. Uryasev and P. M. Pardalos.
22. M. V. Solodov and S. K. Zavriev, Error Stability Properties of Generalized Gradient-Type Algorithms, *J. of Opt. Theory and Applications* **98** (1998) 663–680.
23. P. Tseng, An Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule, *SIAM J. on Optimization* **2** (1998) 506–531.
24. J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms, *IEEE Trans. on Automatic Control* **AC-31** (1986) 803–812.
25. B. Widrow and M. E. Hoff, Adaptive Switching Circuits, *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, part **4** (1960) 96–104.
26. X. Zhao, P. B. Luh, and J. Wang, Surrogate Gradient Algorithm for Lagrangian Relaxation, *J. of Opt. Theory and Applications* **100** (1999) 699–712.