July 28, 1990 LIDS-P-1989

ON THE RATE OF CONVERGENCE OF A PARTIALLY ASYNCHRONOUS GRADIENT PROJECTION ALGORITHM*

by

Paul Tseng†

ABSTRACT

Recently, Bertsekas and Tsitsiklis proposed a partially asynchronous implementation of the gradient projection algorithm of Goldstein and Levitin and Polyak for the problem of minimizing a differentiable function over a closed convex set. In this paper, we analyze the rate of convergence of this algorithm. We show that if the standard assumptions hold (that is, the solution set is nonempty and the gradient of the function is Lipschitz continuous) *and* (i) the isocost surfaces of the objective function, restricted to the solution set, are properly separated and (ii) a certain multifunction associated with the problem is locally upper Lipschitzian, then this algorithm attains a linear rate of convergence.

KEY WORDS. Partially asynchronous computation, gradient projection, locally upper Lipschitzian multifunction, linear convergence.

---

* This research is partially supported by the U.S. Army Research Office, contract DAAL03-86-K-0171 (Center for Intelligent Control Systems) and by the National Science Foundation, grant NSF–DDM–8903385.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Acknowledgement. Thanks are due to Z.–Q. Luo for his many helpful comments on an earlier draft of this paper.

1. Introduction

A frequently encountered problem in optimization concerns finding a stationary point of a continuously differentiable function f in $\Re^m$, the m–dimensional Euclidean space, over a closed convex set $\mathcal{X}$ in $\Re^m$. In other words, it is desired to find a solution to the fixed point problem

$$x = [x - \nabla f(x)]^+,$$

where $[\cdot]^+$ denotes the orthogonal projection operator onto $\mathcal{X}$, i.e., $[x]^+ = \arg\min_{y \in \mathcal{X}} ||x - y||$. In our notation, all vectors are column vectors and $||x||$ denotes the Euclidean norm of x, that is, $||x|| = \sqrt{\langle x, x \rangle}$, where $\langle x, y \rangle$ denotes the Euclidean inner product of x with y.

A well–known iterative method for solving the above problem is the gradient projection algorithm proposed by Goldstein [Gol64] and by Levitin and Polyak [LeP65]. In this algorithm, each new iterate is obtained by moving the previous iterate along the negative gradient direction, and then projecting the resulting point back onto the feasible set $\mathcal{X}$, that is,

$$x := [x - \gamma \nabla f(x)]^+, \tag{1.1}$$

where γ is some appropriately chosen positive stepsize. This algorithm possesses nice numerical properties and has been studied extensively (see [Ber76], [Ber72a], [CaM87], [Che84], [Dun84], [Dun87], [GaB82], [GaB84], [Gol64], [Gol74], [LeP65]).

Recently, Bertsekas and Tsitsiklis [BeT89, Section 7.5] (also see [Tsi84], [TBA86]) proposed a ***partially asynchronous*** implementation of the above algorithm, in which $\mathcal{X}$ is decomposed into the Cartesian product of closed convex sets $\mathcal{X}_1, ..., \mathcal{X}_n$ ($n \geq 1$) and the iteration (1.1) is distributed over n processors, with the i–th processor being responsible for updating the block–component of x belonging to $\mathcal{X}_i$. Each processor carries its own estimate of the solution, communicates to the other processors by message passing, and may act independently of the other processors. Such an "asynchronous" (or "chaotic") computing environment, proposed by Chazan and Miranker [ChM69], offers several advantages over a synchronous (either sequential or parallel) computing environment: for example, the synchronization penalty is low and the fast processors need not wait for the slower ones. In addition, asynchronous computation brings forth interesting and challenging questions about the convergence of algorithms. For a detailed discussion of asynchronous computation, see [BeT89].

It is known, under a standard Lipschiz continuity condition on the gradient ∇f, that if γ is sufficiently small, then every limit point of the iterates generated by the partially asynchronous gradient projection algorithm is a stationary point [BeT89, Sec. 7.5]. However, little is known about the convergence or the rate of convergence of the iterates. In fact, even in the sequential case (i.e., the original gradient projection algorithm), very little is known about the rate of convergence. Rate of convergence analysis typically requires the solution points to be isolated and the objective function f to be locally strongly convex (see [LeP65], [Dun84], [Dun87]), which in general does not hold. Recently, Luo and Tseng [LuT90] (also see

[LuT89], [TsL90a], [TsL90b] for related analyses) proposed a new approach to demonstrating the linear rate of convergence of iterative optimization algorithms, based on bounding the distance to the solution set from a point x near the solution set by the norm of the natural "residual" at x, namely

$$x - [x - \nabla f(x)]^+.$$

Such a local "error bound" does not hold in general, but can be shown to hold for a number of important problem classes, including quadratic programs and strongly convex programs. In this paper, we adapt the approach of Luo and Tseng to analyze the partially asynchronous gradient projection algorithm. In particular, we show that the algorithm attains a linear rate of convergence, assuming only that (i) the solution set is nonempty, (ii) f is bounded from below on $\mathcal{X}$, (iii) ∇f is Lipschitz continuous on $\mathcal{X}$, (iv) the isocost surfaces of the objective function, restricted to the solution set, are "properly separated" from each other, and (v) the above error bound holds near the solution set. Thus, even in the sequential case, our rate of convergence result appears to be a significant improvement over existing ones. [Assumptions (i) to (iv), as we shall see, hold for most problems, so the key assumption is (v).]

This paper proceeds as follows: In Section 2 we describe the partially asychronous gradient projection algorithm and state our main convergence result for this algorithm. In Section 3 we prove the main result for the special case where the computations take place sequentially. In Section 4, we prove the main result, building on the ideas developed in Section 3. In Section 5 we discuss possible extensions of our work.

2. Algorithm Description and Convergence Results

We formally describe the partially asynchronous gradient projection algorithm below (also see [BeT89, Sec. 7.5]). In this algorithm, $\mathcal{X}$ is decomposed into the Cartesian product of closed convex sets $\mathcal{X}_1, ...\mathcal{X}_n$ ($n \geq 1$), that is,

$$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n. \tag{2.1}$$

According to the above product structure of $\mathcal{X}$, let the elements x of $\mathcal{X}$ be decomposed into block–components, so $x = (x_1, x_2, ..., x_n)$, with $x_i \in \mathcal{X}_i$. Let $\nabla_i f(x)$ denote the partial derivative of $f(x)$ with respect to x_i, and let $[x_i]_i^+$ denote the orthogonal projection of x_i onto $\mathcal{X}_i$. Then, for a given *fixed* stepsize $\gamma > 0$, the algorithm generates a sequence of iterates $\{x(1), x(2), ...\}$ in $\mathcal{X}$ according to the formula:

$$x_i(t+1) = \begin{cases} [x_i(t) - \gamma \nabla_i f(x^i(t))]_i^+, & \text{if } t \in T^i; \\ x_i(t), & \text{otherwise,} \end{cases} \qquad i = 1, ..., n, \tag{2.2}$$

where T^i is some subset of $\{0, 1, 2, ...\}$ and $x^i(t)$ is the vector in $\mathcal{X}$ given by

$$x^i(t) = \left(x_1(\tau_1^i(t)), ..., x_n(\tau_n^i(t))\right), \tag{2.3}$$

with each $\tau_j^i(t)$ some nonnegative integer not exceeding t. [The initial iterate $x(0) \in \mathcal{X}$ is assumed given.]

Roughly speaking, T^i is the set of times at which x_i is updated (by processor i); $x^i(t)$ is the solution estimate known to processor i at time t; and $\tau_j^i(t)$ is the time at which the value of x_j used by processor i at time t is generated by processor j (so $t - \tau_j^i(t)$ is effectively the communication delay from processor j to processor i at time t). A key feature of the algorithm is that the components are updated using values which may be out–of–dated.

We make the standing assumption that the iterates are updated in a partially asynchronous manner:

Partial Asynchronism Assumption. There exists an integer $B \geq 1$ such that

(a) $\{t, t+1, ..., t+B-1\} \cap T^i \neq \emptyset$, for all $t \geq 0$ and all i;

(b) $0 \leq t - \tau_j^i(t) \leq B - 1$, for all $t \in T^i$, all j and all i.

[Roughly speaking, the partial asynchronism assumption states that no processor waits an arbitrarily long time to compute or to receive a message from another processor. The justification for this assumption is discussed in Section 7 of [BeT89].]

We make the following standard (and reasonable) assumptions about f and $\mathcal{X}$:

Assumption A.

(a) f is bounded from below on $\mathcal{X}$.

(b) The solution set $\mathcal{X}^* = \{ x \in \Re^m \mid x = [x - \nabla f(x)]^+ \}$ is nonempty.

(c) ∇f is Lipschitz continuous on $\mathcal{X}$, that is,

$$||\nabla f(y) - \nabla f(x)|| \leq L||y - x||, \quad \forall x \in \mathcal{X}, \ \forall y \in \mathcal{X}, \tag{2.4}$$

where $L > 0$ is the Lipschitz constant.

The following convergence result is due to Bertsekas and Tsitsiklis (see Proposition 5.3 in Sec. 7.5 of [BeT89]):

Proposition 2.1. Under Assumption A, there exists a scalar $\gamma_0 > 0$ (depending on L, n and B only) such that if $0 < \gamma < \gamma_0$, then any limit point of the sequence $\{x(t)\}$ generated by the partially asynchronous gradient projection algorithm (2.2), (2.3) is an element of $\mathcal{X}^*$.

The above result is rather weak since it does not assert that $\{x(t)\}$ has a limit point. To prove the convergence of $\{x(t)\}$, we need to make, in addition to Assumption A, the following assumptions on f and $\mathcal{X}$.

Assumption B.

(a) There exists a scalar $\epsilon > 0$ such that

$$x \in \mathcal{X}^*, \quad y \in \mathcal{X}^*, \quad f(x) \neq f(y), \qquad \Rightarrow \qquad ||x - y|| \geq \epsilon.$$

(b) For every η there exist scalars $\delta > 0$ and $\kappa > 0$ such that

$$\phi(x) \leq \kappa||x - [x - \nabla f(x)]^+||, \tag{2.5}$$

for all $x \in \mathcal{X}$ with $f(x) \leq \eta$ and $||x - [x - \nabla f(x)]^+|| \leq \delta$, where we let $\phi(x) = \min_{\bar{x} \in \mathcal{X}^*} ||x - \bar{x}||$.

The main result of this paper is stated below. Its proof, which is quite involved, is given in Section 4.

Proposition 2.2. Under Assumptions A and B, there exists a scalar $\gamma_1 > 0$, depending on L, n, B and $x(0)$ only, such that if $0 < \gamma < \gamma_1$, then the sequence $\{x(t)\}$ generated by the partially asynchronous gradient projection algorithm (2.2)–(2.3) converges at least linearly to an element of $\mathcal{X}^*$ with a B–step convergence ratio of $1 - c\gamma$, where $c > 0$ is some scalar constant.

A few words about Assumption B is in order. Part (a) of Assumption B is a technical assumption which states that the isocost surfaces of f, restricted to the solution set $\mathcal{X}^*$, are "properly separated" from each other. This assumption clearly holds if $\mathcal{X}^*$ is a finite set. More generally, it can be seen to hold if f takes on only a finite number of values on $\mathcal{X}^*$ or if the piecewise–smooth path connected components of $\mathcal{X}^*$ are properly separated from each other. [We say a set is *piecewise–smooth path connected* if any two points in that set can be joined by a piecewise–smooth path lying entirely in that set.] Thus, it holds automatically

when f is convex (since $\mathcal{X}^*$ is then convex) or when $\mathcal{X}$ is polyhedral and f is quadratic (see Lemma 3.1 in [TsL90a]).

Part (b) of Assumption B is closely related to the notion of a locally upper Lipschitzian multifunction (see [Rob81], [Rob82]). More precisely, for any fixed scalar η, let R be the *residual* function given by

$$R(x) = x - [x - \nabla f(x)]^+,$$

restricted to the domain $\{ x \in \mathcal{X} \mid f(x) \leq \eta \}$. Then Assumption B (b) effectively says that the inverse of R, a multifunction, is locally upper Lipschitzian at the origin or, more precisely, there exist scalars $\delta > 0$ and $\kappa > 0$ such that

$$R^{-1}(z) \subseteq R^{-1}(0) + \kappa ||z||^2 \mathcal{B},$$

for all $z \in \Re^m$ with $||z|| \leq \delta$, where $\mathcal{B}$ denotes the unit Euclidean ball in $\Re^m$.

A simple example (e.g., $\mathcal{X} = \Re$ and $f(x) = |x|^\lambda$ with $\lambda > 2$ a fixed scalar) will show that Assumption B (b) does not hold in general. On the other hand, it does hold for a number of important problem classes. For example, it holds when ∇f is strongly monotone and Lipschitz continuous (see [Pan87]). Alternatively, it holds when $\mathcal{X}$ is polyhedral and f is either quadratic (see [Rob81], [TsL90a]) or of the form

$$f(x) = g(Ex) + \langle b, x \rangle,$$

for some $k \times m$ matrix E, some $b \in \Re^m$ and some *strictly* convex twice differentiable function g in $\Re^k$ with $\nabla^2 g$ positive definite everywhere (see [LuT90]). It also holds when $\mathcal{X} = \Re^m$ and f is the dual functional associated with a certain strictly convex network flow problem (see [TsL90b]).

3. Convergence Proof for the Sequential Case

As the proof of Proposition 2.2 is quite intricate, it is instructive to first examine a simpler case to gain a feel for the main ideas used in the proof. In this section, we give a proof of Proposition 2.2 for the special case of the algorithm (2.2)–(2.3) in which $B = 1$ (i.e., the original gradient projection algorithm). We remark that, even for this special case, our convergence result (see Proposition 3.1) appears to be new since it assumes neither convexity of f nor uniqueness of solution (compare with [BeT89, Sec. 3.5.3], [Dun81], [Dun86], [LuT90, Sec. 4], [LeP65]).

For $B = 1$, the partially asynchronous gradient projection algorithm (2.2)–(2.3) reduces to the sequential algorithm

$$x(t+1) = [x(t) - \gamma \nabla f(x(t))]^+, \qquad t = 0, 1, ..., \tag{3.1}$$

with $x(0) \in \mathcal{X}$ given. To analyze the convergence of $\{x(t)\}$ we need the following lemma, which follows from the observation that, for any x and d in $\Re^m$, the function $p(\gamma) = ||x - [x - \gamma d]^+||$ is monotonically increasing in $\gamma > 0$ and the function $p(\gamma)/\gamma$ is monotonically decreasing in $\gamma > 0$ (see Lemma 1 in [GaB84]; also see Lemma 2.2 in [CaM87]).

Lemma 3.1. For any $x \in \mathcal{X}$ and any scalar $\gamma > 0$,

$$\min\{1, \gamma\}||x - [x - \nabla f(x)]^+|| \leq ||x - [x - \gamma \nabla f(x)]^+||.$$

We now state and prove the main result of this section. The proof is patterned after one given in Section 3 of [TsL90a] and is based on using the locally upper Lipschitzian condition (2.5) to show that $\{x(t)\}$ tends toward $\mathcal{X}^*$ [cf. (3.5)] and that, near $\mathcal{X}^*$, the difference in the f value of $x(t+1)$ and that of an element of $\mathcal{X}^*$ nearest to $x(t)$ is at most of the order $||x(t+1) - x(t)||^2$ [see (3.9)].

Proposition 3.1. Under Assumptions A and B, if $0 < \gamma < 2/L$, then the sequence $\{x(t)\}$ generated by the sequential gradient projection algorithm (3.1) converges at least linearly to an element of $\mathcal{X}^*$ with a convergence ratio of $1 - c\gamma$, where $c > 0$ is some scalar constant.

Proof. It is well–known, by using (2.4) and (3.1), that

$$f(x(t+1)) - f(x(t)) \leq -(\frac{1}{\gamma} - \frac{L}{2})||x(t+1) - x(t)||^2, \qquad \forall t. \tag{3.2}$$

(See, for example, [Gol64] or [LeP65].) Since $0 < \gamma < 2/L$ and, by Assumption A (a), f is bounded from below on $\mathcal{X}$, then (3.2) implies

$$x(t) - x(t+1) \to 0, \tag{3.3}$$

so (3.1) and Lemma 3.1 yields $x(t) - [x(t) - \nabla f(x(t))]^+ \to 0$. Since $f(x(t)) \leq f(x(0))$ for all t [cf. (3.2)], this together with Assumption B (b) implies that there exist an index $\bar{t}$ and a scalar $\kappa > 0$ (depending on

$x(0)$) such that, for all $t \geq \bar{t}$, (2.5) holds with $x = x(t)$, so

$$\begin{aligned} ||x(t) - \bar{x}(t)|| &\leq \kappa ||x(t) - [x(t) - \nabla f(x(t))]^+|| \\ &\leq \kappa \max\{1, \frac{1}{\gamma}\} ||x(t) - [x(t) - \gamma \nabla f(x(t))]^+|| \\ &= \kappa \max\{1, \frac{1}{\gamma}\} ||x(t) - x(t+1)||, \end{aligned} \tag{3.4}$$

where $\bar{x}(t)$ denotes an element of $\mathcal{X}^*$ for which $||x(t) - \bar{x}(t)|| = \phi(x(t))$, the second inequality follows from Lemma 3.1, and the equality follows from (3.1). Combining (3.3) with (3.4) gives

$$x(t) - \bar{x}(t) \to 0, \tag{3.5}$$

so $\bar{x}(t) - \bar{x}(t+1) \to 0$. Then, Assumption B (a) implies that $\bar{x}(t)$ eventually settles down at some isocost surface of f, i.e., there exist an index $\hat{t} \geq \bar{t}$ and a scalar $\bar{v}$ such that

$$f(\bar{x}(t)) = \bar{v}, \quad \forall t \geq \hat{t}. \tag{3.6}$$

We have from $\bar{x}(t) \in \mathcal{X}^*$ and $x(t) \in \mathcal{X}$ that $\langle \nabla f(\bar{x}(t)), x(t) - \bar{x}(t) \rangle \geq 0$ and from the Mean Value Theorem that $f(\bar{x}(t)) - f(x(t)) = \langle \nabla f(\psi(t)), \bar{x}(t) - x(t) \rangle$, for some m–vector $\psi(t)$ lying on the line segment joining $\bar{x}(t)$ with $x(t)$. Upon summing these two relations and using (3.6), we obtain

$$\begin{aligned} \bar{v} - f(x(t)) &\leq \langle \nabla f(\psi(t)) - \nabla f(\bar{x}(t)), \bar{x}(t) - x(t) \rangle \\ &\leq ||\nabla f(\psi(t)) - \nabla f(\bar{x}(t))|| ||\bar{x}(t) - x(t)|| \\ &\leq L ||\bar{x}(t) - x(t)||^2, \end{aligned}$$

where the last inequality follows from the Lipschitz condition (2.4) and $||\psi(t) - \bar{x}(t)|| \leq ||x(t) - \bar{x}(t)||$. This together with (3.5) yields

$$\liminf_{t \to \infty} f(x(t)) \geq \bar{v}. \tag{3.7}$$

Since $x(t+1)$ is obtained by projecting $x(t) - \gamma \nabla f(x(t))$ onto $\mathcal{X}$ [cf. (3.1)] and $\bar{x}(t) \in \mathcal{X}$, we have

$$\langle x(t) - \gamma \nabla f(x(t)) - x(t+1), x(t+1) - \bar{x}(t) \rangle \geq 0, \quad \forall t. \tag{3.8}$$

Also, by the Mean Value Theorem, for each $t \geq \hat{t}$ there exists some $\zeta(t)$ lying on the line segment joining $x(t+1)$ with $\bar{x}(t)$ such that

$$f(x(t+1)) - f(\bar{x}(t)) = \langle \nabla f(\zeta(t)), x(t+1) - \bar{x}(t) \rangle,$$

which, when combined with (3.6) and (3.8) yields

$$\begin{aligned} f(x(t+1)) - \bar{v} &= f(x(t+1)) - f(\bar{x}(t)) \\ &= \langle \nabla f(\zeta(t)), x(t+1) - \bar{x}(t) \rangle \end{aligned}$$

$$
\begin{aligned}
&\leq \langle \nabla f(\zeta(t)) - \nabla f(x(t)) + \frac{1}{\gamma}(x(t) - x(t+1)), x(t+1) - \bar{x}(t) \rangle \\
&\leq \left(||\nabla f(\zeta(t)) - \nabla f(x(t))|| + \frac{1}{\gamma}||x(t) - x(t+1)|| \right) ||x(t+1) - \bar{x}(t)|| \\
&\leq \left(L||\zeta(t) - x(t)|| + \frac{1}{\gamma}||x(t) - x(t+1)|| \right) ||x(t+1) - \bar{x}(t)|| \\
&\leq \left(L||x(t+1) - x(t)|| + L||\bar{x}(t) - x(t)|| + \frac{1}{\gamma}||x(t) - x(t+1)|| \right) ||x(t+1) - \bar{x}(t)|| \\
&\leq \left((L + \frac{1}{\gamma})||x(t+1) - x(t)|| + L||x(t) - \bar{x}(t)|| \right) (||x(t+1) - x(t)|| + ||\bar{x}(t) - x(t)||) \\
&\leq \eta_1 ||x(t+1) - x(t)||^2, \qquad (3.9)
\end{aligned}
$$

where the the third inequality follows from the Lipschitz condition (2.4), the fourth inequality follows from the fact that $\zeta(t)$ lies between $x(t+1)$ and $\bar{x}(t)$, and the last inequality follows from (3.4) with η_1 being some scalar constant depending on L, κ and γ only.

Using (3.2) to bound the right hand side of (3.9) gives

$$
f(x(t+1)) - \bar{v} \leq \eta_2 \left(f(x(t)) - f(x(t+1)) \right), \quad \forall t \geq \hat{t},
$$

where η_2 is some positive scalar depending on L, κ and γ only. Upon rearranging terms in the above relation, we obtain

$$
f(x(t+1)) - \bar{v} \leq \frac{\eta_2}{1+\eta_2} \left(f(x(t)) - \bar{v} \right), \quad \forall t \geq \hat{t}.
$$

On the other hand, we have from (3.7) and the fact $f(x(t))$ is monotonically decreasing with t [cf. (3.2)] that $f(x(t)) \geq \bar{v}$ for all t, so the above relation implies that $\{f(x(t))\}$ converges at least linearly to $\bar{v}$. Since $||x(t+1) - x(t)||^2$ is of the order $f(x(t)) - f(x(t+1))$ [cf. (3.2)], this implies that $\{x(t)\}$ converges at least linearly. Since $\phi(x(t)) \to 0$ [cf. (3.5)], then the point to which $\{x(t)\}$ converges is in $\mathcal{X}^*$. That the convergence ratio is of the form $1 - \gamma c$ can be seen by explicitly writing out η_2 as a function of γ. ■

We remark that we need not have assumed γ to be fixed or small in the above analysis, so long as γ is chosen so that $||x(t) - x(t+1)||^2$ is of the order $f(x(t)) - f(x(t+1))$ [cf. (3.2)]. This is an important generalization since, in practice, γ is typically not fixed but determined by some line search rule, such as the Armijo–like rule of Bertsekas [Ber76], and, in this case, the above condition often does hold.

4. Convergence Proof for the General Case

In this section we extend the analysis in Section 3 to prove Proposition 2.2, the main result of this paper. Our argument is very similar in idea to the proof of Proposition 3.1, but, owing to the presence of asynchronism in computations, error quantities arise in many places and have to be carefully estimated. We show that the errors caused by asynchronism are of second order in γ and are negligible when γ is small.

We assume throughout that Assumptions A and B hold. Let $\{x(t)\}$ be a sequence of iterates generated by the partially asynchronous gradient projection algorithm (2.2)–(2.3). For the moment, the only restriction that we place on the stepsize γ is that it be positive. We will, in the course of the proof, impose additional upper bounds on γ.

For each $t \geq 0$, let

$$s_i(t) = x_i(t+1) - x_i(t), \qquad i = 1, ..., n. \tag{4.1}$$

(For notational simplicity, we have defined $s_i(t)$ slightly differently from [BeT89, Sec. 7.5.4].) Then, by (2.2), for every i there holds

$$s_i(t) = \quad 0, \qquad \forall t \notin T^i, \tag{4.2}$$

$$s_i(t) = [x_i(t) - \gamma \nabla_i f(x^i(t))]_i^+ - x_i(t), \qquad \forall t \in T^i. \tag{4.3}$$

For notational simplicity we will use $\Theta(\gamma^k)$, for any integer k, to represent any continuous function $g : (0, \infty) \to \Re$ with the property

$$\lim_{\gamma \downarrow 0} \frac{g(\gamma)}{\gamma^k} = c,$$

for some scalar $c > 0$ depending on L, n, B and $x(0)$ only. We will implicitly assume that γ is always taken sufficiently small so that each $g(\gamma)$ encountered is positive.

First we have the following result analogous to (3.2):

Lemma 4.1.

$$f(x(t+B)) \leq f(x(t)) - \Theta(\gamma^{-1}) \sum_{\tau=t}^{t+B-1} ||s(\tau)||^2 + \Theta(1) \sum_{\tau=t-B}^{t-1} ||s(\tau)||^2, \qquad \forall t \geq 0. \tag{4.4}$$

Proof. For any i and any $t \in T^i$, since $x_i(t) + s_i(t)$ is the orthogonal projection of $x_i(t) - \gamma \nabla_i f(x^i(t))$ onto $\mathcal{X}_i$ [cf. (4.3)] and $x_i(t) \in \mathcal{X}_i$, we have from a well-known property of orthogonal projections that

$$\langle s_i(t), \gamma \nabla_i f(x^i(t)) \rangle \leq -||s_i(t)||^2.$$

Combining this with (4.1)–(4.2) and using (2.4) and an argument analogous to that in [BeT89, pp. 529–530] gives

$$f(x(t+1)) - f(x(t)) \leq -\frac{1 - \gamma L}{\gamma} ||s(t)||^2 + L \sum_{i=1}^{n} \sum_{\tau=t-B}^{t-1} ||s_i(t)|| ||s(\tau)||, \qquad \forall t \geq 0.$$

By using the identity $a \cdot b \leq a^2 + b^2$, we can bound the right hand side of the above relation:

$$\begin{aligned} f(x(t+1)) - f(x(t)) &\leq -\frac{1-\gamma L}{\gamma}||s(t)||^2 + L\sum_{i=1}^{n}\sum_{\tau=t-B}^{t-1}\left(\sqrt{n}||s_i(t)||^2 + \frac{1}{\sqrt{n}}||s(\tau)||^2\right) \\ &= -\frac{1-\gamma L}{\gamma}||s(t)||^2 + L\left(B\sqrt{n}||s(t)||^2 + \sqrt{n}\sum_{\tau=t-B}^{t-1}||s(\tau)||^2\right) \\ &= -\frac{1-\gamma L-\gamma LB\sqrt{n}}{\gamma}||s(t)||^2 + L\sqrt{n}\sum_{\tau=t-B}^{t-1}||s(\tau)||^2. \end{aligned}$$

Applying the above argument successively to $t, t+1, ..., t+B-1$ and we obtain

$$f(x(t+B)) - f(x(t)) \leq -\frac{1-\gamma L-\gamma LB\sqrt{n}}{\gamma}\sum_{\tau=t}^{t+B-1}||s(\tau)||^2 + LB\sqrt{n}\sum_{\tau=t-B}^{t+B-1}||s(\tau)||^2. \tag{4.5}$$

■

[We analyze the B–step decrease in f because, in the worst case, B time units can pass before any component of x is iterated upon.]

By summing (4.4) over all $t = 0, B, 2B, ...$, we see that, for γ sufficiently small so that the $\Theta(\gamma^{-1})$ term dominates the $\Theta(1)$ term in (4.4), there holds

$$\limsup_{t\to\infty} f(x(t)) \leq f(x(0)) - \Theta(\gamma^{-1})\sum_{\tau=0}^{\infty}||s(\tau)||^2.$$

[In fact, it can be seen from (4.5) that it suffices to take $\gamma < L + 3LB\sqrt{n}$.] This implies that $\{f(x(t))\}$ is bounded [cf. Assumption A (a)] and

$$x(t) - x(t+1) \to 0 \tag{4.6}$$

[cf. (4.1)], so, by using (2.4) and (4.3) and the partial asynchronism assumption, we can conclude that

$$x(t) - [x(t) - \gamma\nabla f(x(t))]^+ \to 0. \tag{4.7}$$

(See [BeT89, pp. 530–531] for a more detailed argument.) Up to this point our analysis has followed closely the proof of Proposition 5.1 in [BeT89, Sec. 7.5], but it starts to diverge from here on.

Eq. (4.7) and Lemma 3.1 imply $x(t) - [x(t) - \nabla f(x(t))]^+ \to 0$, and since $\{f(x(t))\}$ is bounded, then, by Assumption B (b), there exists a threshold $\bar{t} \geq 0$ and a scalar $\kappa > 0$ (depending on $x(0)$ only) such that

$$\phi(x(t)) \leq \kappa||x(t) - [x(t) - \nabla f(x(t))]^+||, \quad \forall t \geq \bar{t}.$$

For each t, let $\bar{x}(t)$ be an element of $\mathcal{X}^*$ satisfying $||x(t) - \bar{x}(t)|| = \phi(x(t))$. Then, we have from the above relation and Lemma 3.1 that

$$||x(t) - \bar{x}(t)|| \leq \kappa\max\{1, \frac{1}{\gamma}\}||x(t) - [x(t) - \gamma\nabla f(x(t))]^+||, \quad \forall t \geq \bar{t}. \tag{4.8}$$

Combining (4.7) with (4.8) gives

$$x(t) - \bar{x}(t) \to 0, \tag{4.9}$$

so (4.6) yields $\bar{x}(t) - \bar{x}(t+1) \to 0$. Then, Assumption B (a) implies that $\bar{x}(t)$ eventually settles down at some isocost surface of f, so there exist an index $\hat{t} \geq \bar{t}$ and a scalar $\bar{v}$ such that

$$f(\bar{x}(t)) = \bar{v}, \quad \forall t \geq \hat{t}. \tag{4.10}$$

Then, an argument identical to the proof of (3.7), with (3.5) and (3.6) replaced by (4.9) and (4.10) respectively, gives

$$\liminf_{t\to\infty} f(x(t)) \geq \bar{v}. \tag{4.11}$$

To prove our next main result (Lemma 4.4), we need the following two technical lemmas. The first lemma says that $||x(t) - [x(t) - \gamma\nabla f(x(t))]^+||^2$ is upper bounded by $\sum_{\tau=t}^{t+B-1} ||s(\tau)||^2$ plus a smaller term. The proof of this, trivial in the sequential case [compare with (3.1)], is complicated owing to the presence of asynchronism in the computations.

Lemma 4.2. For all $t \geq 0$ there holds

$$||x(t) - [x(t) - \gamma\nabla f(x(t))]^+||^2 \leq \Theta(1) \sum_{\tau=t}^{t+B-1} ||s(\tau)||^2 + \Theta(\gamma) \sum_{\tau=t-B}^{t-1} ||s(\tau)||^2. \tag{4.12}$$

Proof. Fix any $t \in \{0, 1, ...\}$. For each index $i \in \{1, ..., n\}$ let t^i be the smallest element of T^i that exceeds t. Then [cf. (4.1), (4.2)]

$$x_i(t^i) = x_i(t), \tag{4.13}$$

and, by (4.3),

$$s_i(t^i) = [x_i(t^i) - \gamma\nabla_i f(x^i(t^i))]_i^+ - x_i(t^i). \tag{4.14}$$

Also, by part (a) of the partial asynchronism assumption, there holds $t \leq t^i \leq t + B - 1$. Combining (4.13) with (4.14) and we have

$$\begin{aligned} ||s_i(t^i)|| &= ||[x_i(t) - \gamma\nabla_i f(x^i(t^i))]_i^+ - x_i(t)|| \\ &\geq ||[x_i(t) - \gamma\nabla_i f(x(t))]_i^+ - x_i(t)|| - \gamma||\nabla_i f(x(t)) - \nabla_i f(x^i(t^i))|| \\ &\geq ||[x_i(t) - \gamma\nabla_i f(x(t))]_i^+ - x_i(t)|| - \gamma L||x(t) - x^i(t^i)||, \end{aligned}$$

where the last inequality follows from the Lipschitz condition (2.4). By using the identity $(a - \gamma b)^2 \geq (1-\gamma)a^2 - \gamma(1+\gamma)b^2$, we obtain from the above relation that

$$||s_i(t^i)||^2 \geq (1-\gamma)||[x_i(t) - \gamma\nabla_i f(x(t))]_i^+ - x_i(t)||^2 - \gamma(1+\gamma)L^2||x(t) - x^i(t^i)||^2. \tag{4.15}$$

Also, since $t \leq t^i \leq t+B-1$ so that [by part (b) of the partial asynchronism assumption] $t-B+1 \leq \tau_j^i(t^i) \leq t+B-1$, we have from (4.1) that, for all j,

$$\begin{aligned} ||x_j(t) - x_j(\tau_j^i(t^i))||^2 &\leq \left(\sum_{\tau=t-B+1}^{t+B-1} ||s_j(\tau)|| \right)^2 \\ &\leq 2B \sum_{\tau=t-B+1}^{t+B-1} ||s_j(\tau)||^2, \end{aligned}$$

where the second inequality follows from the identity $(a_1+...+a_{2B})^2 \leq 2B(a_1)^2+...+2B(a_{2B})^2$. Summing the above relation over all j and using (2.3) yields

$$||x(t) - x^i(t^i)||^2 \leq 2B \sum_{\tau=t-B}^{t+B-1} ||s(\tau)||^2.$$

Using the above to bound the right hand side of (4.15) then gives

$$||s_i(t^i)||^2 \geq (1-\gamma)||[x_i(t) - \gamma \nabla_i f(x(t))]_i^+ - x_i(t)||^2 - 2\gamma(1+\gamma)L^2 B \sum_{\tau=t-B}^{t+B-1} ||s(\tau)||^2.$$

Since the choice of i was arbitrary, the above relation holds for all $i \in \{1,...,n\}$, which when summed over all i and using the "obvious" inequality [cf. $t \leq t^i \leq t+B-1$]

$$\sum_{\tau=t}^{t+B-1} ||s(\tau)||^2 = \sum_{i=1}^{n} \sum_{\tau=t}^{t+B-1} ||s_i(\tau)||^2 \geq \sum_{i=1}^{n} ||s_i(t^i)||^2,$$

then gives

$$\sum_{\tau=t}^{t+B-1} ||s(\tau)||^2 \geq (1-\gamma)||[x(t) - \gamma \nabla f(x(t))]^+ - x(t)||^2 - 2\gamma(1+\gamma)L^2 Bn \sum_{\tau=t-B}^{t+B-1} ||s(\tau)||^2.$$

Taking $\gamma < 1$ and rearranging terms in the above relation proves (4.12). ∎

We next have a technical lemma on the behaviour of f over $\mathcal{X}$. Its proof is given in Section 6.

Lemma 4.3. For any x and $x^1,...,x^n$ in $\Re^m$ and any $\bar{x} \in \mathcal{X}$, there holds

$$f(z) - f(\bar{x}) \leq \Theta(\gamma^{-2})||x-\bar{z}||^2 + \Theta(1)\left(||x-\bar{x}||^2 + \sum_{i=1}^{n} ||x-x^i||^2 \right), \tag{4.16}$$

where $\bar{z} = [x - \gamma \nabla f(x)]^+$ and z is the m–vector with components $z_i = [x_i - \gamma \nabla_i f(x^i)]_i^+$.

By using (4.8) and (4.10) together with Lemmas 4.2 and 4.3, we can now upper bound $f(x(t+B)) - \bar{v}$ in a manner analogous to (3.9).

Lemma 4.4. For all $t \geq \hat{t}$ there holds

$$f(x(t+B)) - \bar{v} \leq \Theta(\gamma^{-2}) \sum_{\tau=t}^{t+B-1} ||s(\tau)||^2 + \Theta(\gamma^{-1}) \sum_{\tau=t-B}^{t-1} ||s(\tau)||^2. \tag{4.17}$$

Proof. Fix any $t \geq \hat{t}$. For each i, let t^i denote the smallest element of T^i exceeding t. Then, by (2.2),

$$x_i(t^i + 1) = [x_i(t) - \gamma \nabla_i f(x^i(t^i))]_i^+, \quad \forall i, \tag{4.18}$$

and, by part (a) of the partial asynchronism assumption,

$$t \leq t^i \leq t + B - 1, \quad \forall i. \tag{4.19}$$

Let us apply Lemma 4.3 with $x = x(t)$, $x^i = x^i(t^i)$ for all i, and $\bar{x} = \bar{x}(t)$. This then gives

$$f(z) - f(\bar{x}(t)) \leq \Theta(\gamma^{-2}) r(t) + \Theta(1) \left(||x(t) - \bar{x}(t)||^2 + \sum_{i=1}^n ||x(t) - x^i(t^i)||^2 \right),$$

where z is the m–vector whose i–th component z_i is $x_i(t^i + 1)$ [cf. (4.18)] and for convenience we have let $r(t) = ||x(t) - [x(t) - \gamma \nabla f(x(t))]^+||^2$. By applying (4.8) and (4.10) to the above relation, we obtain the simpler bound

$$\begin{aligned} f(z) - \bar{v} &\leq \Theta(\gamma^{-2}) r(t) + \Theta(1) \sum_{i=1}^n ||x(t) - x^i(t^i)||^2 \\ &= \Theta(\gamma^{-2}) r(t) + \Theta(1) \sum_{i=1}^n \sum_{j=1}^n ||x_j(t) - x_j(\tau_j^i(t^i))||^2, \end{aligned}$$

where the equality follows from (2.3). Since (4.19) holds, then part (b) of the partial asynchronism assumption implies $t - B + 1 \leq \tau_j^i(t^i) \leq t + B - 1$ for all i and all j, so the above relation together with (4.1) yields

$$\begin{aligned} f(z) - \bar{v} &\leq \Theta(\gamma^{-2}) r(t) + \Theta(1) \sum_{j=1}^n \sum_{\tau=t-B+1}^{t+B-1} ||s_j(\tau)||^2 \\ &= \Theta(\gamma^{-2}) r(t) + \Theta(1) \sum_{\tau=t-B+1}^{t+B-1} ||s(\tau)||^2. \end{aligned} \tag{4.20}$$

Also, we have from (4.1), (4.19) and the definition of z_i that

$$x_i(t + B) - z_i = x_i(t + B) - x_i(t^i + 1) = \sum_{\tau=t^i+1}^{t+B-1} s_i(\tau), \quad \forall i,$$

so an argument similar to the proof of (4.4) yields

$$f(x(t + B)) \leq f(z) + \Theta(1) \sum_{\tau=t-B}^{t-1} ||s(\tau)||^2.$$

This combined with (4.20) and then using the definition of $r(t)$ and (4.12) gives (4.17). ∎

By using the bounds (4.4), (4.11) and (4.17), we can now prove the linear convergence of $\{x(t)\}$. To simplify the notation, let

$$\begin{aligned} \alpha(t) &= f(x(t)) - \bar{v}, \\ \beta(t) &= \sum_{\tau=t-B}^{t-1} ||s(\tau)||^2, \end{aligned}$$

for all $t \geq \hat{t}$. Then, we have from (4.4), (4.11) and (4.17) respectively that, for any $t \geq \hat{t}$,

$$\alpha(t+B) \leq \alpha(t) - \gamma^{-1} A_1 \beta(t+B) + A_2 \beta(t), \tag{4.21}$$

$$0 \leq \liminf_{\tau \to \infty} \alpha(\tau), \tag{4.22}$$

$$\alpha(t+B) \leq \gamma^{-2} A_3 \beta(t+B) + \gamma^{-1} A_3 \beta(t), \tag{4.23}$$

where A_1, A_2, A_3 are positive scalars depending on L, n, B and $x(0)$ only. [Of course, all of this come under the implicit assumption that γ is taken sufficiently small.] Notice that we are now explicitly writing out the constant in the $\Theta(\cdot)$ notation. For this part of the proof, the constant matters. Our goal will be to show, by induction, that $\{\alpha(t)\}$ and, in particular, $\{\beta(t)\}$ converge at least linearly (see Lemma 4.6). [Notice that $\beta(t) \geq 0$ but, in contrast to the sequential case, $\alpha(t)$ may be negative. This fortunately does not complicate our proof to any significant degree.]

Fix any $t \geq \hat{t} + B$. Applying (4.23) to bound the $\beta(t+B)$ term in (4.21) and then rearranging terms gives

$$(1 + \gamma A_1/A_3)\alpha(t+B) \leq \alpha(t) + (A_1 + A_2)\beta(t).$$

Also, by substituting $t - B$ for t in (4.21) and rearranging terms, we obtain $\gamma^{-1} A_1 \beta(t) \leq \alpha(t-B) - \alpha(t) + A_2 \beta(t-B)$, which when applied to bound the right hand side of the above relation yields

$$(1 + \gamma A_1/A_3)\alpha(t+B) \leq \alpha(t) + \gamma(1 + A_2/A_1)\left(\alpha(t-B) - \alpha(t) + A_2\beta(t-B)\right).$$

After rearranging terms, we obtain

$$\alpha(t+B) \leq \frac{1}{1+\gamma A_1/A_3}\left((1-\gamma A_4)\alpha(t) + \gamma A_4\left(\alpha(t-B) + A_2\beta(t-B)\right)\right), \tag{4.24}$$

where for convenience we let $A_4 = 1 + A_2/A_1$. Also, for any $k \geq 2$, we have from repeated applications of (4.21) that

$$\begin{aligned}
\alpha(t+kB) &\leq \alpha(t) - \gamma^{-1} A_1 \sum_{l=1}^{k} \beta(t+lB) + A_2 \sum_{l=0}^{k-1} \beta(t+lB) \\
&= \alpha(t) - (\gamma^{-1} A_1 - A_2) \sum_{l=1}^{k-1} \beta(t+lB) - \gamma^{-1} A_1 \beta(t+kB) + A_2 \beta(t).
\end{aligned}$$

By taking $\gamma < A_1/A_2$, we then obtain from the nonnegativity of $\beta(\tau)$, for all τ, that

$$\alpha(t+kB) \leq \alpha(t) - (\gamma^{-1} A_1 - A_2)\beta(t+B) + A_2 \beta(t).$$

By letting $k \to \infty$ in the above relation, we obtain from (4.22) that

$$0 \leq \alpha(t) - (\gamma^{-1} A_1 - A_2)\beta(t+B) + A_2 \beta(t),$$

which upon rearranging terms gives

$$\beta(t+B) \le \frac{\gamma}{A_1 - \gamma A_2}(\alpha(t) + A_2\beta(t)). \tag{4.25}$$

Fix any two scalars $a > 0$ and $b > 0$ satisfying

$$8A_3A_4A_2b = A_1a, \tag{4.26}$$

with a and b taken sufficiently large so that

$$\alpha(\hat{t}) \le a, \qquad \alpha(\hat{t}+B) \le a, \qquad \beta(\hat{t}) \le b, \qquad \beta(\hat{t}+B) \le b. \tag{4.27}$$

Also let

$$c = \frac{A_1}{2A_3 + 2A_1}. \tag{4.28}$$

By using (4.24)–(4.28), we obtain the following main result of this section:

Lemma 4.6. There exists a scalar $\gamma_1 > 0$ (depending on A_1 up to A_4 only) such that if $0 < \gamma < \gamma_1$, then, for all $r = 0, 1, 2, ...$, there holds

$$\alpha(\hat{t} + rB) \le a\rho^{r-1}, \tag{4.29}$$

$$\beta(\hat{t} + rB) \le b\rho^{r-1}, \tag{4.30}$$

where

$$\rho = 1 - \gamma c. \tag{4.31}$$

Proof. The proof is by induction on r. By (4.27), both (4.29) and (4.30) hold for $r = 0, 1$. Suppose that (4.29) and (4.30) hold for all r from 0 up to some $k \ge 1$. We show below that if $0 < \gamma < \gamma_1$, for some γ_1 depending on A_1 up to A_4 only, then (4.29) and (4.30) hold for $r = k + 1$. This would then complete the induction on r and show that, if $0 < \gamma < \gamma_1$, then (4.29) and (4.30) hold for all $r \ge 0$. For convenience, we denote $t = \hat{t} + kB$ in what follows.

First we show that

$$\alpha(t+B) \le a\rho^k. \tag{4.32}$$

Since (4.29) and (4.30) hold for all r up to k, we obtain from (4.24) and by taking $\gamma < 1/A_4$ that

$$\begin{aligned}
\alpha(t+B) &\le \frac{1}{1+\gamma A_1/A_3}\left((1-\gamma A_4)a\rho^{k-1} + \gamma A_4\left(a\rho^{k-2} + A_2b\rho^{k-2}\right)\right) \\
&\le \frac{1}{1+\gamma A_1/A_3}\left(1 - \gamma A_4 + \gamma A_4(1 + A_2b/a)(1+\gamma 2c)\right)a\rho^{k-1} \\
&= \frac{1}{1+\gamma A_1/A_3}\left(1 + \gamma^2 2A_4c + \gamma A_4(1+\gamma 2c)A_2b/a\right)a\rho^{k-1} \\
&\le \frac{1}{1+\gamma A_1/A_3}\left(1 + \gamma A_1/(2A_3)\right)a\rho^{k-1} \\
&= \left(1 - \frac{\gamma A_1}{2A_3 + \gamma 2A_1}\right)a\rho^{k-1} \\
&\le \left(1 - \frac{\gamma A_1}{2A_3 + 2A_1}\right)a\rho^{k-1},
\end{aligned}$$

where the second inequality follows from taking $\gamma \leq 1/(2c)$ and using the bound $\rho^{-1} \leq 1 + 2c\gamma$ [cf. (4.31)], the third inequality follows from (4.26) and $\gamma \leq 1/(2c)$ and by taking $\gamma \leq A_1/(8A_3A_4c)$, and the last inequality follows by taking $\gamma < 1$. The above relation together with (4.28) and (4.31) yields (4.32).

Next we show that

$$\beta(t+B) \leq b\rho^k. \tag{4.33}$$

Since (4.29) and (4.30) hold for all r from 0 up to k, we have from (4.25) that

$$\begin{aligned}\beta(t+B) &\leq \frac{\gamma}{A_1 - \gamma A_2}(a\rho^{k-1} + A_2 b\rho^{k-1}) \\ &= \frac{\gamma(a/b + A_2)}{A_1 - \gamma A_2} b\rho^{k-1}.\end{aligned}$$

By taking γ sufficiently small so $\gamma(a/b + A_2) \leq (A_1 - \gamma A_2)(1 - \gamma c)$, we obtain from (4.31) that (4.33) holds.

Since (4.32) and (4.33) hold, then (4.29) and (4.30) hold for $r = k + 1$. ■

Lemma 4.6 implies that $\{\beta(t)\}$ converges at least linearly with a B–step convergence ratio of $1 - \gamma c$. Since $||x(t) - x(t-B)||^2 \leq B\beta(t)$ for all t [cf. (4.1) and definition of $\beta(t)$], this shows that $\{x(t)\}$ converges at least linearly with a B–step convergence ratio of $\sqrt{1 - \gamma c}$, which is at most $1 - \gamma c/2$. Since $\phi(x(t)) \to 0$ [cf. (4.9)], it follows that the point to which $\{x(t)\}$ converges is an element of $\mathcal{X}^*$.

5. Extensions

For simplicity we have assumed that ∇f is Lipschitz continuous everywhere on $\mathcal{X}$ [cf. (2.4)], but this need not be so. More generally, it suffices that f tends to ∞ at any boundary point of its effective domain and that ∇f is Lipschitz continuous on each level set of f, intersected with $\mathcal{X}$.

In [TsB86] (also see Section 7.6 of [BeT89]) is discussed a distributed asynchronous routing algorithm. This algorithm is based on the idea of gradient projection and, by making suitable modifications to our analysis, it is possible to show that, under conditions analogous to Assumptions A and B, this algorithm also attains a linear rate of convergence.

A drawback of our main result (Proposition 2.2) is that convergence requires the algorithm to take very small steps. Intuitively, if f is approximately separable with respect to the components x_i (that is, $f(x) \approx \sum_i f_i(x_i)$ for some functions f_i), then the algorithm should be able to take much larger steps. This notion can be made precise by incorporating the effect of second order quantities such as $\frac{\partial f}{\partial x_i \partial x_j}$ (assuming that f is twice differentiable) into the convergence analysis.

6. Proof of Lemma 4.3

For each i, since z_i is the orthogonal projection of $x_i - \gamma\nabla_i f(x^i)$ onto the closed convex set $\mathcal{X}_i$ and [cf. $\bar{x} \in \mathcal{X}$ and (2.1)] $\bar{x}_i \in \mathcal{X}_i$, we have

$$\langle z_i - \bar{x}_i, x_i - \gamma\nabla_i f(x^i) - z_i \rangle \geq 0.$$

Also, by the Mean Value Theorem, there exists some ζ lying on the line segment joining z with $\bar{x}$ such that $f(z) - f(\bar{x}) = \langle z - \bar{x}, \nabla f(\zeta) \rangle$, so the above relation yields

$$\begin{aligned} f(z) - f(\bar{x}) &= \langle z - \bar{x}, \nabla f(\zeta) \rangle \\ &\leq \sum_{i=1}^{n} \langle z_i - \bar{x}_i, \nabla_i f(\zeta) - \nabla_i f(x^i) + \frac{1}{\gamma}(x_i - z_i) \rangle \\ &\leq \sum_{i=1}^{n} ||z_i - \bar{x}_i|| \left(||\nabla_i f(\zeta) - \nabla_i f(x^i)|| + \frac{1}{\gamma}||x_i - z_i|| \right) \\ &\leq \sum_{i=1}^{n} ||z_i - \bar{x}_i|| \left(L||\zeta - x^i|| + \frac{1}{\gamma}||x_i - z_i|| \right) \\ &\leq \sum_{i=1}^{n} ||z - \bar{x}|| \left(L||\zeta - x^i|| + \frac{1}{\gamma}||x - z|| \right), \end{aligned} \tag{6.1}$$

where the third inequality follows from the Lipschitz condition (2.4). Now we bound the right hand side of (6.1). Since ζ is between z and $\bar{x}$, we have $||\zeta - x|| \leq ||z - x|| + ||\bar{x} - x||$ so that

$$\begin{aligned} ||\zeta - x^i|| &\leq ||\zeta - x|| + ||x - x^i|| \\ &\leq ||z - x|| + ||\bar{x} - x|| + ||x - x^i||. \end{aligned} \tag{6.2}$$

Also, we have from the definition of z and $\bar{z}$ and the nonexpansive property of the projection operators $[\cdot]_i^+$, $i = 1, ..., n$, that

$$\begin{aligned} ||z - \bar{z}||^2 &= \sum_{i=1}^{n} ||z_i - \bar{z}_i||^2 \\ &= \sum_{i=1}^{n} ||[x_i - \gamma\nabla_i f(x^i)]_i^+ - [x_i - \gamma\nabla_i f(x)]_i^+||^2 \\ &\leq \sum_{i=1}^{n} \gamma^2 ||\nabla_i f(x^i) - \nabla_i f(x)||^2 \\ &\leq \sum_{i=1}^{n} \gamma^2 L^2 ||x^i - x||^2, \end{aligned} \tag{6.3}$$

where the last inquality follows from the Lipschitz condition (2.4).

By bounding the right hand side of (6.1) using (6.2) and the triangle inequality, we have

$$
\begin{aligned}
f(z) - f(\bar{x}) &\leq \sum_{i=1}^{n} ||z - \bar{x}|| \Big(L||\bar{x} - x|| + L||x - x^i|| + (L + \frac{1}{\gamma})||z - x|| \Big) \\
&\leq \sum_{i=1}^{n} (||z - \bar{z}|| + ||\bar{z} - x|| + ||x - \bar{x}||) \Big(L||\bar{x} - x|| + L||x - x^i|| \\
&+ (L + \frac{1}{\gamma})(||z - \bar{z}|| + ||\bar{z} - x||) \Big) \\
&\leq n(3(L + \frac{1}{\gamma})^2 + 4)(||z - \bar{z}||^2 + ||\bar{z} - x||^2) + n(3L^2 + 4)||x - \bar{x}||^2 + 3L^2 \sum_{i=1}^{n} ||x - x^i||^2,
\end{aligned}
$$

where the last inequality follows from expanding out the product in the line above and then using the bound $a \cdot b \leq a^2 + b^2$ on each term of the expansion. Using (6.3) to bound the $||z - \bar{z}||^2$ term in the above relation and we obtain (4.16).

References

[Ber76] Bertsekas, D. P., On the Goldstein–Levitin–Polyak Gradient Projection Method, *IEEE Trans. Automat. Control*, 21 (1976), 174–184.

[Ber82a] Bertsekas, D. P., Projected Newton Methods for Optimization Problems with Simple Constraints, *SIAM J. Contr. & Optim.*, 20 (1982), 221–246.

[Ber82b] Bertsekas, D. P., *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, NY (1982).

[BeT89] Bertsekas, D. P. and Tsitsiklis, J. N., *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ (1989).

[CaM87] Calamai, P. H. and Moré, J. J., Projected Gradient Methods for Linearly Constrained Problems, *Math. Prog.*, 39 (1987), 93–116.

[ChM69] Chazan, D., and Miranker, W. L., Chaotic Relaxation, *Lin. Algeb. & Appl.*, 2 (1969), 199–222.

[Che84] Cheng, Y. C., On The Gradient–Projection Method for Solving the Nonsymmetric Linear Complementarity Problem, *J. Appl. Math. and Optim.*, 43 (1984), 527–540.

[Dun81] Dunn, J. C., Global and Asymptotic Convergence Rate Estimates for a Class of Projected Gradient Processes, *SIAM J. Contr. & Optim.*, 19 (1981), 368–400.

[Dun86] Dunn, J. C., On the Convergence of Projected Gradient Processes to Singular Critical Points, *J. Optim. Theory & Applic.*, 55 (1987), 203–216.

[GaB82] Gafni, E. M., and Bertsekas, D. P., Convergence of a Gradient Projection Method, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Report No. P–1201, Cambridge, MA (1982).

[GaB84] Gafni, E. M., and Bertsekas, D. P., Two–Metric Projection Methods for Constrained Optimization, *SIAM J. Contr. & Optim.*, 22 (1984), 936–964.

[Gol64] Goldstein, A. A., Convex Programming in Hilbert Space, *Bull. Am. Math. Soc.*, 70 (1964), 709–710.

[Gol74] Goldstein, A. A., On Gradient Projection, *Proc. 12–th Ann. Allerton Conference on Circuits and Systems*, Allerton Park, Ill. (1974), 38–40.

[LeP65] Levitin, E. S. and Poljak, B. T., Constrained Minimization Methods, *Z. Vycisl. Mat. i Mat. Fiz.*, 6 (1965), 787-823. English translation in *USSR Comput. Math. Phys.*, 6 (1965), 1–50.

[LuT89] Luo, Z.-Q. and Tseng, P., On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Report No. P–1924, Cambridge, MA (December 1989; revised July 1990); to appear in *J. Optim. Theory & Appl.*

[LuT90] Luo, Z.-Q. and Tseng, P., On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Report No. P–1979, Cambridge, MA (June 1990).

[Mor89] Moré, J. J., Gradient Projection Techniques for Large–Scale Optimization Problems, *Proc. of the 28–th Conference on Decision and Control*, Tampa, Florida (December 1989).

[Pan85] Pang, J.-S., A Posteriori Error Bounds for the Linearly–Constrained Variational Inequality Problem, *Math. Oper.*, 12 (1985), 474–484.

[Rob81] Robinson, S. M., Some Continuity Properties of Polyhedral Multifunctions, *Math. Prog. Study*, 14 (1981), 206–214.

[Rob82] Robinson, S. M., Generalized Equations and Their Solutions, Part II: Applications to Nonlinear Programming, *Math. Prog. Study*, 14 (1982), 200–221.

[TsL90a] Tseng, P., and Luo, Z.-Q., Error Bound and Convergence Analysis of Matrix Splitting Algorithms for the Affine Variational Inequality Problem, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Report P-1988, Cambridge, MA (June 1990).

[TsL90b] Tseng, P., and Luo, Z.-Q., On the Linear Convergence of the Relaxation Method for Nonlinear Network Flow Problems, in preparation.

[Tsi84] Tsitsiklis, J. N., Problems in Decentralized Decision Making and Computation, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA (1984).

[TsB86] Tsitsiklis, J. N., and Bertsekas, D. P., Distributed Asynchronous Optimal Routing in Data Networks, *IEEE Trans. Automat. Contr.* AC-31, 325–332.

[TBA86] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms, *IEEE Trans. Automat. Contr.* AC-31, 803–812 (1986).