The Asynchronous PALM Algorithm for Nonsmooth Nonconvex Problems

Damek Davis

April 2, 2016

Abstract We introduce the Asynchronous PALM algorithm, a new extension of the Proximal Alternating Linearized Minimization (PALM) algorithm for solving nonsmooth, nonconvex optimization problems. Like the PALM algorithm, each step of the Asynchronous PALM algorithm updates a single block of coordinates; but unlike the PALM algorithm, the Asynchronous PALM algorithm eliminates the need for sequential updates that occur one after the other. Instead, our new algorithm allows each of the coordinate blocks to be updated asynchronously and in any order, which means that any number of computing cores can compute updates in parallel without synchronizing their computations. In practice, this asynchronization strategy often leads to speedups that increase linearly with the number of computing cores.

We introduce two variants of the Asynchronous PALM algorithm, one stochastic and one deterministic. In the stochastic *and* deterministic cases, we show that cluster points of the algorithm are stationary points. In the deterministic case, we show that the algorithm converges globally whenever the Kurdyka-Łojasiewicz property holds for a function closely related to the objective function, and we derive its convergence rate in a common special case. Finally, we provide a concrete case in which our assumptions hold.

1 Introduction

In this paper, we tackle the nonsmooth nonconvex optimization problem

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ f(x_1,\ldots,x_m)+\sum_{i=1}^{m} r_j(x_j), \tag{1.1}$$

where $\mathcal{H}$ is a finite dimensional Euclidean space, f is a C^1 function, and each r_j is a proper, lower semi-continuous function. Our approach is similar to the Proximal Alternating Linearized Minimization (PALM)

This material is based upon work supported by the National Science Foundation under Award No. 1502405.

D. Davis
Department of Mathematics, University of California, Los Angeles
Los Angeles, CA 90025, USA
E-mail: damek@math.ucla.edu

algorithm [5], which repeatedly, in a cyclic order, runs through coordinate blocks and performs prox-gradient steps: for all $k \in \mathbb{N}$, get x^{k+1} from x^k via

$$\text{For } j = 1, \ldots, m$$
$$x_j^{k+1} \in \operatorname*{arg\,min}_{x_j \in \mathcal{H}_j} \left\{ r_j(x_j) + \langle \nabla_j f(x_1^{k+1}, \ldots, x_{j-1}^{k+1}, x_j^k, \ldots, x_m^k), x_j - x_j^k \rangle + \frac{1}{2\gamma_j^k} \|x_j - x_j^k\|^2 \right\}.$$

We, too, perform alternating prox-gradient steps on (1.1), but we differ from PALM in two respects: (a) we allow both stochastic and deterministic block update orders, and (b) we break the synchronization enforced by PALM by allowing several computing cores to work in parallel on local prox-gradient updates which are then chaotically, and without coordination, written to a global shared memory source. The theoretical difficulty and practical importance of (a) is negligible, but without it we could not perform (b), which is theoretically new for the PALM algorithm and sometimes results in big practical improvements for other algorithms [21,19,22,23,18,17]. We expect similar improvements to result from Asynchronous PALM, but for a wider class of problems that includes matrix factorization and Generalized Low Rank Models (GLRM) [27].

Like most recent work on first order algorithms for nonsmooth, nonconvex optimization [1,9,28,6,8,12, 13,6,7,15,14], our analysis relies on the *nonsmooth Kurdyka-Łojasiewicz* (KL) property [2], which relates the growth of a function to growth of its subgradients. And we also follow the general proof recipe given in the original PALM paper [5]. But we are eventually forced to depart from the theory in the PALM paper because asynchronous parallel updates introduce errors, so we must analyze a decreasing *Lyapunov function* that absorbs the errors, rather than the not necessarily decreasing objective function $f + \sum_j r_j$. That becomes the main theoretical contribution of this paper, and more generally, it presents a first step for designing asynchronous parallel first-order algorithms for solving nonsmooth, nonconvex optimization problems more complex than (1.1).

2 Notation

The Asynchronous PALM algorithm solves (1.1) in a finite dimensional Hilbert space, like $\mathbb{R}^n$, which we call $\mathcal{H}$. We assume the Hilbert space $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$ is a product of $m \in \mathbb{N}$ other Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_m$; we also define $\mathcal{H}_{-j} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_{j-1} \times \mathcal{H}_{j+1} \times \cdots \times \mathcal{H}_m$. Given a vector $x \in \mathcal{H}$, we denote its jth component by $x_j \in \mathcal{H}_j$. Given a sequence $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{H}$ and a vector $h \in \mathbb{N}^m$, we define

$$(\forall k \in \mathbb{N}) \qquad x^{k-h} = (x_1^{k-h_1}, \ldots, x_m^{k-h_m}) \tag{2.1}$$

and use the convention that $x_j^k = x_j^0$ if $k \leq 0$; we also let $\mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$ denote the set of cluster points of $\{x^k\}_{k\in\mathbb{N}}$. For $j \in \{1, \ldots, m\}$, we let $\langle \cdot, \cdot \rangle : \mathcal{H}_j \times \mathcal{H}_j \to \mathbb{R}$ denote the inner product on $\mathcal{H}_j$, and we let $\|\cdot\|$ be the corresponding norm (i.e., we do not distinguish between the different norms on the components of $\mathcal{H}$). For all $x, y \in \mathcal{H}$, we let $\langle x, y \rangle = \sum_{j=1}^m \langle x_j, y_j \rangle$ and $\|x\| := \sqrt{\langle x, x \rangle}$ be the standard inner product and norm on $\mathcal{H}$. A *box* $B := B_1 \times \cdots \times B_m \subseteq \mathcal{H}$ is any product of balls $B_j \subseteq \mathcal{H}_j$.

We define

$$r(x) := \sum_{j=1}^m r_j(x) \qquad \text{and} \qquad \Psi := f + r$$

Throughout this paper, we assume that Ψ is bounded below and that r is prox-bounded [25, Definition 1.23]:

$$(\exists \lambda_r > 0) : (\forall x \in \mathcal{H}), (\forall \lambda \leq \lambda_r) \qquad \mathbf{prox}_{\lambda r}(x) := \operatorname*{arg\,min}_{y \in \mathcal{H}} \{r(y) + \frac{1}{2\lambda}\|x - y\|^2\} \neq \emptyset.$$

For any point $x \in \mathcal{H}$, we denote by $x_{-j} \in \mathcal{H}_{-j}$, the point x with the jth component removed. With this notation, we assume that

$$(\forall j \in \{1, \ldots, m\}), (\forall x \in \mathcal{H}) \qquad \nabla_j f(x_{-j}; \cdot) : \mathcal{H}_j \to \mathcal{H}_j \text{ is } L_j(x_{-j})\text{-Lipschitz.}$$

In particular, we always have the descent lemma [20]:

$$(\forall j \in \{1, \ldots, m\}), (\forall x \in \mathcal{H}), (\forall y \in \mathcal{H}_j)$$
$$f(x_{-j}; x_j) \leq f(x_{-j}; y) + \langle x_j - y, \nabla_j f(x_{-j}; y)\rangle + \frac{L_j(x_{-j})}{2}\|x_j - y\|^2.$$

We also assume that for any bounded set B, there exists a constant L_B such that

$$\nabla f : \mathcal{H} \to \mathcal{H} \text{ is } L_B\text{-Lipschitz continuous on } B,$$

which is guaranteed if, for example, f is C^2.

For any proper, lower semi-continuous function $g : \mathcal{H} \to (-\infty, \infty]$, we let $\partial_L g : \mathcal{H} \to 2^{\mathcal{H}}$ denote the *limiting subdifferential* of g; see [25, Definition 8.3].

For any $\eta \in (0, \infty)$, we let F_η denote the class of concave continuous functions $\varphi : [0, \eta) \to \mathbb{R}_+$ for which $\varphi(0) = 0$; φ is C^1 on $(0, \eta)$ and continuous at 0; and for all $s \in (0, \eta)$, we have $\varphi'(s) > 0$.

A function $g : \mathcal{H} \to (-\infty, \infty]$ has the *Kurdyka-Łojasiewicz* (KL) property at $\overline{u} \in \operatorname{dom}(\partial_L g)$ provided that there exists $\eta \in (0, \infty)$, a neighborhood U of $\overline{u}$, and a function $\varphi \in F_\eta$ such that

$$(\forall u \in U \cap \{u' \mid g(\overline{u}) < g(u') < g(\overline{u}) + \eta\}) \qquad \varphi'(g(u) - g(\overline{u}))\operatorname{dist}(0, \partial_L g(u)) \geq 1.$$

The function g is said to be a *KL function* provided it has the KL property at each point $u \in \operatorname{dom}(g)$.

We work with an underlying probability space denoted by $(\Omega, \mathcal{F}, P)$, and we assume that the space $\mathcal{H}$ is equipped with the Borel σ-algebra. We always let $\sigma(X) \subseteq \mathcal{F}$ denote the sub σ-algebra generated by a random variable or vector X. We use the shorthand a. s. to denote almost sure convergence of a sequence of random variables.

Most of the concepts that we use in this paper can be found in [3, 25].

3 The Algorithms and Assumptions

In this paper, we study the behavior of two different algorithms, one stochastic and one deterministic. Both algorithms solve (1.1). They differ only in one respect: how the active coordinates are selected at each iteration. In the stochastic case, larger stepsizes are allowed, but at the cost of weaker convergence guarantees, namely solely subsequence convergence. In the deterministic case, only smaller stepsizes are allowed, but by leveraging the nonsmooth Kurdyka-Łojasiewicz property, global sequence convergence is guaranteed—provided the sequence of iterates is bounded.

Besides increased flexibility in choosing active coordinates, the significant difference between the proposed algorithms and the standard PALM algorithm lies in the not necessarily cyclic update order and the delays, which are conveniently summarized by vectors of integers:

$$d_k \in \{0, \ldots, \tau\}^m.$$

In both the stochastic and deterministic cases, the gradient of f is evaluated at x^{k-d_k} (see (2.1) for the definition of this vector). In general, $x^{k-d_k} \notin \{x^k, x^{k-1}, \ldots, x^{k-\tau}\}$, but we always have $x_j^{k-d_{k,j}} \in \{x_j^k, x_j^{k-1}, \ldots, x_j^{k-\tau}\}$.

These choices make for a practical delay model. For example, in a software implementation of the algorithms we introduce below, we might (1) read the inconsistent iterate x^{k-d_k}, (2) evaluate the partial gradient $\nabla_j f(x^{k-d_k})$, (3) read the current iterate x_j^k, which might have changed from $x_j^{k-d_{k,j}}$ while we were busy computing the gradient, (4) evaluate the proximal mapping $\mathbf{prox}_{\gamma_j^k r_j}(x_j^k - \gamma_j^k \nabla_j f(x^{k-d_k}))$, and finally, (5) write any element of this proximal mapping to the computer memory.

The two algorithms follow:

Algorithm 1 (Stochastic Asynchronous PALM) *Choose $x^0 \in \mathcal{H}$, $c \in (0,1)$, and $M > 0$. Then for all $k \in \mathbb{N}$, perform the following three steps:*

1. *Sample $j_k \in \{1, \ldots, m\}$ uniformly at random.*
2. *Choose $\gamma_j^k = \min\left\{c\left(L_j(x_{-j}^{k-d_k}) + \frac{2M\tau}{m^{1/2}}\right)^{-1}, \lambda_r\right\}$.*
3. *Set*

$$x_j^{k+1} \in \begin{cases} \arg\min_{x_j \in \mathcal{H}_j}\left\{r_j(x_j) + \langle \nabla_j f(x^{k-d_k}), x_j - x_j^k\rangle + \frac{1}{2\gamma_j^k}\|x_j - x_j^k\|^2\right\} & \text{if } j = j_k; \\ \{x_j^k\} & \text{otherwise.} \end{cases}$$

□

Assumption 1 (Stochastic Asynchronous PALM) 1. *For all $j \in \{1, \ldots, m\}$, the mapping $L_j : \mathcal{H}_{-j} \to \mathbb{R}$ is measurable.*

2. *For all $j \in \{1, \ldots, m\}$, there is a measurable selection $\zeta_j : \mathcal{H} \times (0, \infty) \to \mathcal{H}$ of the set-valued mapping $\mathbf{prox}_{(\cdot) r_j}(\cdot)$ such that for all $k \in \mathbb{N}$, we have*

$$x_{j_k}^{k+1} = \zeta_{j_k}\left(x_{j_k}^k - \gamma_{j_k}^k \nabla_{j_k} f(x^{k-d_k}), \gamma_{j_k}^k\right)$$

(which by Part 1 of this assumption makes $x_{j_k}^{k+1}$ $\sigma(x^1, \ldots, x^k)$-measurable).

3. *There exists $L > 0$ such that for all $k \in \mathbb{N}$ and $j \in \{1, \ldots, m\}$, we have*

$$\sup_{k \in \mathbb{N}}\{L_j(x_{-j}^{k-d_k})\} \leq L. \qquad \text{a. s.}$$

4. *For all $k \in \mathbb{N}$, the constant M satisfies*

$$\|\nabla f(x^k) - \nabla f(x^{k-d_k})\| \leq M\|x^k - x^{k-d_k}\| \qquad \text{a. s. .}$$

In the deterministic case, we have complete control over the indices j_k. Thus, it is often the case that the quantity

$$\rho_\tau := \sup_{j,k} |\{h \mid k - \tau \leq h \leq \tau, j_h = j\}|,$$

which always satisfies $\rho_\tau \leq \tau$, is actually substantially less than τ. In fact, ρ_τ is only equal to τ in the extreme case in which j_h is constant for τ consecutive values of h. However, for the ideal case in which $\tau = O(m)$, in which we have $O(m)$ independent processors, all of which are equally powerful, and in which each coordinate prox-gradient subproblem is equally easy or difficult to solve, we expect that $\rho_\tau = O(1)$. Thus, in the following algorithm, we replace τ by $\sqrt{\rho_\tau \tau}$ in the stepsize formula.

Algorithm 2 (Deterministic Asynchronous PALM) *Choose $x^0 \in \mathcal{H}$, $c \in (0,1)$, and $M > 0$. Then for all $k \in \mathbb{N}$, perform the following three steps:*

1. *Choose $j_k \in \{1, \ldots, m\}$.*
2. *Choose* $\gamma_j^k = \min\left\{c\left(L_j(x_{-j}^{k-d_k}) + 2M\sqrt{\rho_\tau \tau}\right)^{-1}, \lambda_r\right\}$.
3. *Set*

$$x_j^{k+1} \in \begin{cases} \arg\min_{x_j \in \mathcal{H}_j} \left\{r_j(x_j) + \langle \nabla_j f(x^{k-d_k}), x_j - x_j^k\rangle + \frac{1}{2\gamma_j^k}\|x_j - x_j^k\|^2\right\} & \text{if } j = j_k; \\ \{x_j^k\} & \text{otherwise.} \end{cases}$$

□

Assumption 2 (Deterministic Asynchronous PALM) 1. *There exists $K \in \mathbb{N}$ such that for all $k \in \mathbb{N}$, we have $\{1, \ldots, m\} \subseteq \{j_{k+1}, \ldots, j_{k+K}\}$.*
2. *There exists $L > 0$ such that for all $k \in \mathbb{N}$ and $j \in \{1, \ldots, m\}$, we have*

$$\sup_{k \in \mathbb{N}}\{L_j(x_{-j}^{k-d_k})\} \leq L.$$

3. *For all $k \in \mathbb{N}$, the constant M satisfies*

$$\|\nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k})\| \leq M\|x^k - x^{k-d_k}\|.$$

3.1 A Convergence Theorem for the Semi-Algebraic Case

Semi-Algebraic functions are an important class of objectives for which the Asynchronous PALM algorithm converges:

Definition 3.1 (Semi-Algebraic Functions) A function $\Psi : \mathcal{H} \to (0, \infty]$ is a *semi-algebraic* provided that $\mathrm{gra}(\Psi) = \{(x, \Psi(x)) \mid x \in \mathcal{H}\}$ is a semi-algebraic set, which in turn means that there exists a finite number of real polynomials $g_{ij}, h_{ij} : \mathcal{H} \times \mathbb{R} \to \mathbb{R}$ such that

$$\mathrm{gra}(\Psi) := \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{u \in \mathcal{H} \mid g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0\}.$$

Not only does Algorithm 2 converge when Ψ is semi-algebraic, but we can also determine how quickly it converges.

Theorem 3.1 (Global Convergence of Deterministic Asynchronous PALM) *Suppose that Ψ is coercive, semi-algebraic, and ∇f is M-Lipschitz continuous on the minimal box B containing the level set $\{x \mid \Psi(x) \leq \Psi(x^0)\}$. Then $\{x^k\}_{k\in\mathbb{N}}$ from Algorithm 2 globally converges to a stationary point x of Ψ.*

Moreover, $\Psi(x^k)-\Psi(x)$ either converges in a finite number of steps, linearly, or sublinearly, and it always converges at a rate no worse than $o((k+1)^{-1})$, depending on a certain exponent of semi-algebraic functions.

This theorem follows from Theorems 6.1 and 5.4, proved below.

3.2 General Convergence Results

In Sections 4 and 5, we show that (provided $\{x^k\}_{k\in\mathbb{N}}$ is bounded) the cluster points of Algorithms 1 and 2 (a. s.) converge to stationary points Ψ; we also show that the objective value $\Psi(x^k)$ (a. s.) converges, that its limit exists (a. s.) and, if x^0 is not a stationary point of Ψ, that the limit is less than $\Psi(x^0)$ (in expectation). This is the content of Theorems 4.1 and 5.1

The strongest guarantee we have for the stochastic case (Algorithm 1) is that cluster points are stationary points; this is because there is no obvious deterministic Lyapunov function that we can apply the KL inequality to, only a supermartingale. In contrast, the strongest guarantee for the deterministic case (Algorithm 2) is that $\{x^k\}_{k\in\mathbb{N}}$ globally converges whenever it is bounded *and* a Lyapunov function is a KL function. This is the content of Theorem 5.3.

3.3 The Strengths of Assumptions 1 and 2

Unless the best possible coordinate Lipschitz constant is chosen, i.e., unless

$$L_j(x_{-j}) = \sup_{y,y'\in\mathcal{H}_j; y\neq y'} \frac{\|\nabla_j f(x_{-j};y) - \nabla_j f(x_{-j};y')\|}{\|y-y'\|}, \tag{3.1}$$

we cannot guarantee that $L_j(x_{-j})$ is measurable; however, as

$$g(y,y',x_{-j}) := \|\nabla_j f(x_{-j};y) - \nabla_j f(x_{-j};y')\|\|y-y'\|^{-1}$$

is continuous on the open set $\mathcal{H}_j^2 \times \mathcal{H}_{-j}\backslash\{(y,y',x_{-j}) \mid y = y'\}$, it follows that L_j defined as in (3.1) is indeed measurable.

The existence of a measurable selection of $\mathbf{prox}_{(\cdot)r_j}(\cdot)$ is not a strong assumption; as shown in [25, Exercise 14.38], it follows from our assumptions on r.

Part 3 of Assumption 1 and Part 2 of Assumption 2 hold as long as the sequence $\{x^{k-d_k}\}_{k\in\mathbb{N}}$ is (a. s.) bounded.

Part 4 of Assumption 1 and Part 3 of Assumption 2 are strong but can certainly be ensured, for example, by either of two obvious sufficient conditions: (a) ∇f is globally Lipschitz continuous with constant M or (b) each regularizer r_j has a bounded domain, in which case $\{x^k\}_{k\in\mathbb{N}} \cup \{x^{k-d_k}\}_{k\in\mathbb{N}}$ is a bounded set, and because ∇f is Lipschitz on bounded sets, this effectively enforces (a) where it need be true. (Notice, too, that Part 4 of Assumption 1 is stronger than Part 3 of Assumption 2 because the left hand side of the bound depends only on the j_kth partial derivative. In particular, for sparsely coupled problems, the M in Part 3 of Assumption 2 can be substantially smaller than the M in Part 4 of Assumption 1.)

3.4 What's New for Asynchronous Optimization Algorithms?

Asynchronous optimization algorithms are typically identical to synchronous optimization algorithms except that they use delayed information wherever possible. Usually, these delays are not imposed by the user and occur naturally because multiple processors are chaotically updating the coordinates of a vector of real numbers whenever they finish an assigned task, for example, the task may be a prox-gradient update as in Algorithms 1 and 2. Abstractly, most asynchronous algorithms take the form of a mapping $T : \mathcal{H} \times \mathcal{H} \times \{1, \ldots, m\} \to \mathcal{H}$

$$x^{k+1} := T(x^k, x^{k-d_k}, j_k),$$

which constructs the next iterate from the current iterate by blending current and stale information, and our algorithms are no different.

Asynchronous optimization algorithms differ most sharply, then, in the choice of T. In the past, T has taken a few different forms, for example, for $\gamma > 0$,[1]

$$T(x^k, x^{k-d_k}, j_k) = \begin{cases} (T^0(x^{k-d_k}))_j & \text{if } j = j_k \\ x_j^k & \text{otherwise.} \end{cases} \quad \text{for a Lipschitz map } T^0 : \mathcal{H} \to \mathcal{H};$$

$$T(x^k, x^{k-d_k}, j_k) = \begin{cases} \mathrm{Proj}_{X_j}(x_j^k - \gamma \nabla_j f(x^{k-d_k})) & \text{if } j = j_k \\ x_j^k & \text{otherwise.} \end{cases} \quad \text{for a convex } X_j \subseteq \mathcal{H}_j \text{ and } (C^1)\ f : \mathcal{H} \to \mathbb{R};$$

$$T(x^k, x^{k-d_k}, j_k) = \begin{cases} x_j^k - \gamma(S(x^{k-d_k}))_j & \text{if } j = j_k \\ x_j^k & \text{otherwise.} \end{cases} \quad \text{for a cocoercive map } S : \mathcal{H} \to \mathcal{H}.$$

All three types of maps appear in the classic textbook [4]. But since that book was released (over 25 years ago), several of the assumptions on these mappings have been weakened; see, for example, the works in [21, 19, 22, 23, 18, 17, 16, 26, 10], some of which contain stochastic variants or nonconvex extensions of the above mappings. (Readers interested in the history of asynchronous algorithms should see [22, Section 1.5].)

The innovation of Algorithms 1 and 2 compared to prior work is twofold: (i) we include the nonsmooth, nonconvex function r and (ii) we use the nonsmooth KL property to guarantee *global convergence* of x^k to a stationary point. For example, the second of the three above asynchronous mappings falls within our assumptions, except that in our case the sets X_j need not be convex—a feature not previously available in asynchronous algorithms (in this case $r_j = \iota_{X_j}$ is an indicator function, which is the prototypical example of a nonsmooth, nonconvex function).

4 The Stochastic Case

In this section, we analyze Algorithm 1, which allows for large stepsizes, but requires stricter problem assumptions than Algorithm 2 does for obtaining subsequence convergence. To make any guarantees in the stochastic case, it is best to assume that r has bounded domain, which implies that $\{x^k\}_{k \in \mathbb{N}}$ is bounded.

We advise the reader that

Assumption 1 is in effect throughout this section.

Theorem 4.1 (Convergence in the Stochastic Case) *Let $\mathcal{F}_k = \sigma(x^1, \ldots, x^k)$ and let $\mathcal{G}_k = \sigma(j_k)$. Assume that $\{j_k\}_{k \in \mathbb{N}}$ are IID and for all $k \in \mathbb{N}$, $\{\mathcal{F}_k, \mathcal{G}_k\}$ are independent.*

[1] Let $\beta > 0$. A map $S : \mathcal{H} \to \mathcal{H}$ is called β-cocoercive if, for all $x, y \in \mathcal{H}$, we have $\beta \|S(x) - S(y)\|^2 \leq \langle S(x) - S(y), x - y \rangle$.

Then, if $\{x^k\}_{k\in\mathbb{N}}$ is almost surely bounded, the following hold: there exists a subset $\Omega_1 \subseteq \Omega$ with measure 1 such that, for all $\omega \in \Omega_1$,

1. *$\mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$ is nonempty and is contained in the set of stationary points of Ψ.*
2. *The objective function Ψ is finite and constant on $\mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$. In addition, the objective values $\Psi(x^k(\omega))$ converge, and if x^0 is not a stationary point of Ψ, then*

$$\mathbb{E}\left[\lim_{k\to\infty} \Psi(x^k)\right] < \Psi(x^0).$$

Proof **Notation.** We define a few often repeated items:

1. **Full update vector.** For all $k \in \mathbb{N}$ and $j \in \{1,\dots,m\}$, we let
$$w_j^k = \zeta_j\left(x_j^k - \gamma_j^k \nabla_j f(x^{k-d_k}), \gamma_j^k\right)$$
and define $w^k := (w_1^k,\dots,w_m^k)$. The random vector w^k is $\mathcal{F}_k$ measurable by Assumption 1.
2. **The Lyapunov function.** Define a function $\Phi : \mathcal{H}^{1+\tau} \to \mathcal{H}^{1+\tau}$:
$$(\forall\, x(0), x(1), \dots, x(\tau) \in \mathcal{H})$$
$$\Phi(x(0), x(1), \dots, x(\tau)) = f(x(0)) + r(x(0)) + \frac{M}{2\sqrt{m}} \sum_{h=1}^{\tau} (\tau - h + 1)\|x(h) - x(h-1)\|^2.$$
3. **The last time an update occured.** We let $l(k,j) \in \mathbb{N}$ be the last time coordinate j was updated:
$$l(k,j) = \max(\{q \mid j_q = j_k, q < k\} \cup \{0\}).$$

Part 1: Two essential elements feature in our proof. The indispensable *supermartingale convergence theorem* [24, Theorem 1], with which we show that a pivotal sequence of random variables converges, is our hammer for nailing down the effect of randomness in Algorithm 1:

Theorem 4.2 (Supermartingale convergence theorem) *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathfrak{F} := \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be an increasing sequence of sub σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$. Let $b \in \mathbb{R}$, let $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ be sequences of $[b,\infty)$ and $[0,\infty)$-valued random variables, respectively, such that for all $k \in \mathbb{N}$, X_k and Y_k are $\mathcal{F}_k$-measurable and*

$$(\forall k \in \mathbb{N}) \qquad \mathbb{E}\left[X_{k+1} \mid \mathcal{F}_k\right] + Y_k \le X_k.$$

Then $\sum_{k=0}^{\infty} Y_k < \infty$ a. s. and X_k a. s. converges to a $[b,\infty)$-valued random variable.

The other equally indispensable element of our proof is the next inequality, which, when taken together with the supermartingale convergence theorem, will ultimately show that Algorithm 1 converges: with

$$X_k := \Phi(x^k, x^{k-1}, \dots, x^{k-\tau}) \qquad \text{and} \qquad Y_k := \frac{1}{2m}\sum_{j=1}^{m}\left(\frac{1}{\gamma_j^k} - L_j(x_{-j}^k) - \frac{2M\tau}{m^{1/2}}\right)\|w_j^k - x_j^k\|^2,$$

the supermartingale inequality holds

$$(\forall k \in \mathbb{N}) \qquad \mathbb{E}\left[X_{k+1} \mid \mathcal{F}_k\right] + Y_k \le X_k. \tag{4.1}$$

So, by the supermartingale convergence theorem, the sequence Y_k is a. s. summable and X_k a. s. converges to a $[\inf_{x\in\mathcal{H}} \Psi, \infty)$-valued random variable X^*.

At this point, we can conclude that several limits exist:

1. Because $\sum_{k=0}^{\infty} Y_k < \infty$ a. s., we conclude that $\|w_j^k - x_j^k\|$ a. s. converges to 0.
2. Because $\|x_j^{k+1} - x_j^k\| \le \|w_j^k - x_j^k\|$, we conclude that $\|x_j^{k+1} - x_j^k\|$ a. s. converges to 0.
3. Because $\|x^{k+1} - x^k\|$ a. s. converges to 0, we conclude that, for any fixed $l \in \mathbb{N}$, both terms $\|x^{k-l} - x^{k-l-1}\|$ and $\|x^k - x^{k-d_k}\|$ a. s. converge to 0.
4. Because for any fixed $l \in \mathbb{N}$, $\|x^{k-l} - x^{k-l-1}\|$ a. s. converges to zero and because X_k a. s. converges to an $\mathbb{R}$-valued-random variable X^*, we conclude that $f(x^k) + r(x^k)$ a. s. converges X^*.

These limits imply that certain subgradients of $f + r$ a. s. converge to zero; namely, if

$$(\forall j) \qquad A_j^k := \frac{1}{\gamma_j^k}(x_j^k - w_j^k) + \nabla_j f(w^k) - \nabla_j f(x^{k-d_k}),$$

then a quick look at optimality conditions verifies that $(A_1^k, \dots, A_m^k) \in \nabla f(w^k) + \partial_L r(w^k) = \partial_L (f + r)(w^k)$, and moreover,

$$\|(A_1^k, \dots, A_m^k)\| \le \max_{j,k} \left\{ \frac{1}{\gamma_j^k} \right\} \|x^k - w^k\| + \|\nabla f(w^k) - \nabla f(x^{k-d_k})\| \to 0 \text{ a. s.}.$$

These limits also imply that all cluster points of $\{x^k\}_{k\in\mathbb{N}}$ are a. s. stationary points—provided there is a full measure set Ω_1 such that for all $\omega \in \Omega_1$ and for every converging subsequence $w^{k_q}(\omega) \to x$ we have $f(w^{k_q}(\omega)) + r(w^{k_q}(\omega)) \to f(x) + r(x)$.

To show this, let's fix a set of full measure $\Omega_1 \subseteq \Omega$ such that for all $\omega \in \Omega_1$, the sequence $\{x^k(\omega)\}$ is bounded and all of the above limits hold. Because $\|x^k(\omega) - w^k(\omega)\| \to 0$, the cluster point sets $\mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$ and $\mathcal{C}(\{w^k(\omega)\}_{k\in\mathbb{N}})$ are equal. Thus, if we fix a cluster point $x \in \mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$, say $x^{k_q}(\omega) \to x$, then $w^{k_q}(\omega) \to x$. Similarly, we have

$$x^{k_q - d_{k_q}}(\omega) \to x \qquad \text{and} \qquad \lim_{q\to\infty} f(x^{k_q}(\omega)) = \lim_{q\to\infty} f(w^{k_q}(\omega)) = f(x).$$

Proving that $\lim_{q\to\infty} r_j(x_j^{k_q}(\omega)) = \lim_{q\to\infty} r_j(w_j^{k_q}(\omega)) = r_j(x_j)$ is a little subtler because r_j is not continuous; it is merely lower semicontinuous.

In what follows, we suppress the dependence of $l(k, j)$ on ω and assume that for our particular choice of ω, $l(k, j) \to \infty$ as $k \to \infty$; if $l(k, j)$ is eventually constant, then $x_j^k(\omega)$ is eventually constant, and then the limits claimed for $r_j(x_j^{k_q}(\omega))$ hold at once.

First,

$$r_j(w_j^{k_q}(\omega)) \le r_j(x_j^{k_q}(\omega)) - \langle \nabla_j f(x^{k_q - d_{k_q}}(\omega)), x_j^{k_q}(\omega) - w_j^{k_q}(\omega) \rangle - \frac{1}{2\gamma_j^k} \|w_j^{k_q}(\omega) - x_j^{k_q}(\omega)\|^2.$$

So $\liminf_{q\to\infty} \left(r_j(w_j^{k_q}(\omega)) - r_j(x_j^{k_q}(\omega)) \right) \le 0$ because $x_j^{k_q}(\omega) - w_j^{k_q}(\omega) \to 0$ and $\nabla f(x^{k_q - d_{k_q}}(\omega))$ is bounded as $q \to \infty$.

Second, for k_q large enough that $l(k_q, j) > 0$ and for any $y \in \mathcal{H}_j$, we have

$$\begin{aligned} r_j(x_j^{k_q}(\omega)) + \langle \nabla_j f(x^{l(k_q,j) - d_{l(k_q,j)}}(\omega)), x_j^{k_q}(\omega) - x_j^{l(k_q,j)}(\omega) \rangle + \frac{1}{2\gamma_j^{l(k_q,j)}} \|x_j^{k_q}(\omega) - x_j^{l(k_q,j)}(\omega)\|^2 \\ \le r_j(y) + \langle \nabla_j f(x^{l(k_q,j) - d_{l(k_q,j)}}(\omega)), y - x_j^{l(k_q,j)}(\omega) \rangle + \frac{1}{2\gamma_j^{l(k_q,j)}} \|y - x_j^{l(k_q,j)}(\omega)\|^2. \end{aligned}$$

This inequality becomes useful after rearranging, setting $y = w_j^{k_q}(\omega)$, and applying the cosine rule to break up the difference of norms:

$$\begin{aligned} r_j(x_j^{k_q}(\omega)) &\leq r_j(w_j^{k_q}(\omega)) + \langle \nabla_j f(x^{l(k_q,j)-d_{l(k_q,j)}}(\omega)), w_j^{k_q}(\omega) - x_j^{k_q}(\omega)\rangle \\ &\quad + \frac{1}{2\gamma_j^{l(k_q,j)}}\left[\|w_j^{k_q}(\omega) - x_j^{l(k_q,j)}(\omega)\|^2 - \|x_j^{k_q}(\omega) - x_j^{l(k_q,j)}(\omega)\|^2\right] \\ &= r_j(w_j^{k_q}(\omega)) + \langle \nabla_j f(x^{l(k_q,j)-d_{l(k_q,j)}}(\omega)), w_j^{k_q}(\omega) - x_j^{k_q}(\omega)\rangle \\ &\quad + \frac{1}{2\gamma_j^{l(k_q,j)}}\left[2\langle w_j^{k_q}(\omega) - x_j^{l(k_q,j)}(\omega), w_j^{k_q}(\omega) - x_j^{k_q}(\omega)\rangle - \|w_j^{k_q}(\omega) - x_j^{k_q}(\omega)\|_j^2\right]. \end{aligned}$$

All the iterates are assumed to be bounded, the inverse step sizes, $(2\gamma_j^{l(k_q,j)})^{-1}$, are bounded, $\nabla_j f$ is continuous, and $w_j^k(\omega) - x_j^k(\omega) \to 0$, so we have

$$\liminf_{q\to\infty}\left(r_j(x_j^{k_q}(\omega)) - r_j(w_j^{k_q}(\omega))\right) \leq 0.$$

Altogether,

$$\lim_{q\to\infty}\left(r_j(x_j^{k_q}(\omega)) - r_j(w_j^{k_q}(\omega))\right) = 0.$$

This difference converges to zero, but we still need to show that the sequence of objective values $r_j(x_j^{k_q}(\omega))$ converges to $r_j(x_j)$.

For this, we use two properties: First, by lower semicontinuity, we have

$$\liminf_{q\to\infty} r_j(x_j^{k_q}(\omega)) \geq r_j(x_j).$$

Second, by the definition of $w_j^{k_q}(\omega)$ as a proximal point, we have

$$\begin{aligned} &\limsup_{q\to\infty} r_j(w_j^{k_q}(\omega)) \\ &\leq \limsup_{q\to\infty}\left(r_j(x_j) + \langle \nabla_j f(x^{k_q-d_{k_q}}(\omega)), x_j - w_j^{k_q}(\omega)\rangle + \frac{1}{2\gamma_j^k}\|x_j - x_j^{k_q}(\omega)\|_j^2.\right) \\ &\leq r_j(x_j). \end{aligned}$$

Therefore, $\lim_{q\to\infty} r_j(x_j^{k_q}(\omega)) = \lim_{q\to\infty} r_j(w_j^{k_q}(\omega)) = r_j(x_j)$, and altogether,

$$f(w^{k_q}(\omega)) + r(w^{k_q}(\omega)) \to f(x) + r(x); \qquad (A_1^k, \ldots, A_m^k) \to 0;$$

and, hence, $0 \in \partial_L(f+r)(x)$.

We finish the proof of Part 1 with the proof of (4.1).

Proof (of (4.1)) We bound the smooth term first:

$$\mathbb{E}\left[f(x^{k+1}) \mid \mathcal{F}_k\right] \leq \frac{1}{m}\left(\sum_{j=1}^{m} f(x^k) + \langle \nabla_j f(x^k), w_j^k - x_j^k\rangle + \frac{L_j(x_{-j}^k)}{2}\|w_j^k - x_j^k\|^2\right)$$

Next we bound the nonsmooth term:

$$\mathbb{E}\left[r_j(x_j^{k+1}) \mid \mathcal{F}_k\right] \leq r_j^k(x_j^k) - \frac{1}{m}\langle \nabla_j f(x^{k-d_k}), w_j^k - x_j^k\rangle - \frac{1}{2\gamma_j^k m}\|w_j^k - x_j^k\|^2.$$

Both terms together now:

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) + \sum_{j=1}^{m} r_j(x_j^{k+1}) \mid \mathcal{F}_k\right] \leq f(x^k) + \sum_{j=1}^{m} r_j(x_j^k) + \frac{1}{m}\langle \nabla f(x^k) - \nabla f(x^{k-d_k}), w^k - x^k\rangle \\ - \frac{1}{2m}\sum_{j=1}^{m}\left(\frac{1}{\gamma_j^k} - L_j(x_{-j}^k)\right)\|w_j^k - x_j^k\|^2.\end{aligned}$$

The cross term needs care. In particular, the following sequence of inequalities is true for any $C > 0$:

$$\begin{aligned}&\langle \nabla f(x^k) - \nabla f(x^{k-d_k}), w^k - x^k\rangle \\ &\leq M\|x^k - x^{k-d_k}\|\|w^k - x^k\| \qquad \text{(by Assumption 1)} \\ &\leq \frac{M^2}{2C}\|x^k - x^{k-d_k}\|^2 + \frac{C}{2}\|w^k - x^k\|^2 \\ &\leq \frac{M^2}{2C}\sum_{j=1}^{m} d_{k,j}\sum_{h=k-d_{k,j}+1}^{k}\|x_j^h - x_j^{h-1}\|^2 + \frac{C}{2}\|w^k - x^k\|^2 \qquad \text{(by Jensen's inequality)} \\ &\leq \frac{M^2\tau}{2C}\sum_{j=1}^{m}\sum_{h=k-\tau+1}^{k}\|x_j^h - x_j^{h-1}\|^2 + \frac{C}{2}\|w^k - x^k\|^2 \\ &= \left(\frac{M^2\tau}{2C}\sum_{h=k-\tau+1}^{k}(h-k+\tau)\|x^h - x^{h-1}\|^2 - \frac{M^2\tau}{2C}\sum_{h=k-\tau+2}^{k+1}(h-(k+1)+\tau)\|x^h - x^{h-1}\|^2\right) \\ &\quad + \frac{M^2\tau^2}{2C}\|x^{k+1} - x^k\|^2 + \frac{C}{2}\|w^k - x^k\|^2.\end{aligned}$$

We collect all the alternating terms in the sequence $\{\kappa_k\}_{k\in\mathbb{N}}$, defined by

$$\kappa_k := \frac{M}{2\sqrt{m}}\sum_{h=k-\tau+1}^{k}(h-k+\tau)\|x^h - x^{h-1}\|^2,$$

and set $C = M\tau(\sqrt{m})^{-1}$. Thus, from $\mathbb{E}\left[\|x^{k+1} - x^k\|^2 \mid \mathcal{F}_k\right] = m^{-1}\|w^k - x^k\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\kappa_{k+1} \mid \mathcal{F}_k\right] \\
&\leq \kappa_k - \frac{1}{m}\langle \nabla f(x^k) - \nabla f(x^{k-d_k}), w^k - x^k\rangle + \frac{M^2\tau^2}{2mC}\mathbb{E}\left[\|x^{k+1} - x^k\|^2 \mid \mathcal{F}_k\right] + \frac{C}{2m}\|w^k - x^k\|^2 \\
&= \kappa_k - \frac{1}{m}\langle \nabla f(x^k) - \nabla f(x^{k-d_k}), w^k - x^k\rangle + \left(\frac{M^2\tau^2}{2m^2C} + \frac{C}{2m}\right)\|w^k - x^k\|^2 \\
&= \kappa_k - \frac{1}{m}\langle \nabla f(x^k) - \nabla f(x^{k-d_k}), w^k - x^k\rangle + \frac{M\tau}{m^{3/2}}\|w^k - x^k\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[f(x^{k+1}) + \sum_{j=1}^m r_j(x_j^{k+1}) + \kappa_{k+1} \mid \mathcal{F}_k\right] \\
&\leq f(x^k) + \sum_{j=1}^m r_j(x_j^k) + \kappa_k - \frac{1}{2m}\sum_{j=1}^m \left(\frac{1}{\gamma_j^k} - L_j(x_{-j}^k) - \frac{2M\tau}{m^{1/2}}\right)\|w_j^k - x_j^k\|^2. \qquad (4.2)
\end{aligned}$$

In particular for all $k \in \mathbb{N}$, we have $\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) = f(x^k) + \sum_{j=1}^m r_j(x_j^k) + \kappa_k$, so (4.1) follows. □

Part 2: Let $\omega \in \Omega_1$ (where Ω_1 is defined in Part 1), let C denote the limit of $\Psi(x^k(\omega))$ as $k \to \infty$ (which exists by Part 1), let $x \in \mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$, and suppose that $x^{k_q}(\omega) \to x$. Then $C = \lim_{q\to\infty}\Psi(x^{k_q}(\omega)) = \Psi(x)$ (again, by Part 1). Thus, Ψ is constant on $\mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$.

The bound on the limit of the objective value is a consequence of (4.1): First,

$$\Phi(x^0, x^{-1}, \ldots, x^{-\tau}) = \Psi(x^0).$$

Second, by the tower property of expectations

$$\begin{aligned}
(\forall k \in \mathbb{N}) \qquad &\mathbb{E}\left[\Phi(x^{k+1}, x^k, \ldots, x^{k-\tau+1})\right] \leq \mathbb{E}\left[\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) - Y_k\right]; \\
\implies &\mathbb{E}\left[\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau})\right] \leq \Psi(x^0) - \sum_{i=0}^{k-1}\mathbb{E}\left[Y_i\right].
\end{aligned}$$

Third, Fatou's lemma implies that

$$\mathbb{E}\left[\liminf_{k\to\infty}\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau})\right] \leq \liminf_{k\to\infty}\mathbb{E}\left[\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau})\right] \leq \Psi(x^0) - \sum_{i=0}^{\infty}\mathbb{E}\left[Y_i\right]. \qquad (4.3)$$

We leverage this bound and Part 1, which shows that that $\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau})$ a. s. converges and $\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) - \Psi(x^k) \to 0$ a. s., to show that

$$\mathbb{E}\left[\lim_{k\to\infty}\Phi(x^k, x^{k-1}, \ldots, x^{k-\tau})\right] = \mathbb{E}\left[\lim_{k\to\infty}\Psi(x^k)\right].$$

Finally, we have only the strict decrease property left to prove: If x^0 is not a stationary point, then $\mathbb{E}[Y_0] = \mathbb{E}\left[\|w^0 - x^0\|\right] = \|w^0 - x^0\| > 0$. Thus, the decrease property follows from (4.3). □

Remark 4.1 The connectedness and compactness of $\mathcal{C}(\{x^k(\omega)\}_{k\in\mathbb{N}})$ are implied by the limit $x^k(\omega) - x^{k+1}(\omega) \to 0$; see [5, Remark 3.3] for details.

5 The Deterministic Case

Stochastic Asynchronous PALM (Algorithm 1) allows for large stepsizes, but stochasticity makes it difficult to show that the sequence of points $\{x^k\}_{k\in\mathbb{N}}$ actually converges, so we do not pursue such a result. Instead in this section we prove that the sequence of points $\{x^k\}_{k\in\mathbb{N}}$ generated by Deterministic Asynchronous PALM (Algorithm 2) converges, but at the cost of using a smaller stepsize.

The key property for us here, but unavailable in the stochastic setting, is the KL property, which we will assume holds for a function $\Phi : \mathcal{H}^{1+\tau} \to \mathcal{H}^{1+\tau}$, defined by

$$(\forall\, x(0), x(1), \ldots, x(\tau) \in \mathcal{H})$$
$$\Phi(x(0), x(1), \ldots, x(\tau)) = f(x(0)) + r(x(0)) + \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}} \sum_{h=1}^{\tau} (\tau - h + 1)\|x(h) - x(h-1)\|^2.$$

Then we proceed in two parts: first, we show that the cluster points, if any, of the sequence

$$z^k := (x^k, \ldots, x^{k-\tau})$$

are of the form $(x, \ldots, x)$ for some $x \in \mathcal{H}$ *and* x and $(x, \ldots, x)$ are stationary points of Ψ and Φ respectively; and second, we show that if Φ is a KL function and if the sequence z^k is bounded, it will converge, i.e., it has only one cluster point. Along the way we will see that if x^0 is not a stationary point, Algorithm 2 decreases the objective value below that of $\Psi(x^0)$.

We advise the reader that

Assumption 2 is in effect throughout this section.

5.1 Cluster points

Theorem 5.1 (Convergence in the Deterministic Case) *The sequence $\{x^k\}$ lies completely within the level set:*

$$\{x^k\}_{k\in\mathbb{N}} \subseteq \{x \mid \Psi(x) \leq \Psi(x^0)\}$$

Moreover if $\{x^k\}_{k\in\mathbb{N}}$ is bounded, then

1. *The set $\mathcal{C}(\{z^k\}_{k\in\mathbb{N}})$ (respectively $\mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$) is nonempty and contained in the set of stationary points of Φ (respectively Ψ). Moreover,*

$$\mathcal{C}(\{z^k\}_{k\in\mathbb{N}}) = \{(x, \ldots, x) \in \mathcal{H}^{1+\tau} \mid x \in \mathcal{C}(\{x^k\}_{k\in\mathbb{N}})\}$$

2. *The objective function Φ (respectively Ψ) is finite and constant on $\mathcal{C}(\{z^k\}_{k\in\mathbb{N}})$ (respectively $\mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$). In addition, the objective values $\Phi(z^k)$ (respectively $\Psi(x^k)$) converge, and if x^0 is not a stationary point of Ψ, then*

$$\big(\forall x^* \in \mathcal{C}(\{x^k\}_{k\in\mathbb{N}})\big) \qquad \Psi(x^*) = \lim_{k\to\infty} \Psi(x^k) < \Psi(x^0).$$

Proof **Notation.** We let $l(k,j) \in \mathbb{N}$ be the last time coordinate j was updated:

$$l(k,j) = \max(\{q \mid j_q = j, q < k\} \cup \{0\}).$$

We delay the proof of the level set inclusion for a moment and return to it at the end of the proof.

Part 1: This proof is similar to the stochastic proof, but has the added simplicity of being completely deterministic. For example, we show that with

$$X_k := \Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) \qquad \text{and} \qquad Y_k := \left(\frac{1}{\gamma_{j_k}^k} - L_{j_k}(x_{-j_k}^k) - 2M\sqrt{\rho_\tau \tau}\right) \|x_{j_k}^{k+1} - x_{j_k}^k\|^2,$$

the Fejér inequality holds

$$(\forall k \in \mathbb{N}) \qquad X_{k+1} + Y_k \leq X_k, \tag{5.1}$$

which implies that $\sum_{k=0}^{\infty} Y_k < \infty$ and that X_k converges to a real number X^* (X_k is lower bounded); and with this inequality in hand, we have

1. Because $\sum_{k=0}^{\infty} Y_k < \infty$, we conclude that $\|x_{j_k}^{k+1} - x_{j_k}^k\|$ converges to 0.
2. Because $\|x^{k+1} - x^k\| \leq \|x_{j_k}^{k+1} - x_{j_k}^k\|$, we conclude that $\|x^{k+1} - x^k\|$ converges to 0.
3. Because $\|x^{k+1} - x^k\|$ converges to 0, we conclude that, for any fixed $l \in \mathbb{N}$, all three terms $\|x^{k-l} - x^{k-l-1}\|$, $\|x^k - x^{k-d_k}\|$, and $\|x^k - x^{l(k,j)-d_{l(k,j)}}\|$ converge to 0.
4. Because for any fixed $l \in \mathbb{N}$, $\|x^{k-l} - x^{k-l-1}\|$ converges to zero and because X_k converges to X^*, we conclude that $f(x^k) + r(x^k)$ converges X^*.

These limits imply that certain subgradients of Φ converge to zero; namely, if, for all k and j, we set

$$A_j^k = \begin{cases} \frac{1}{\gamma_j^k}(x_j^k - x_j^{k+1}) + \nabla_j f(x^{k+1}) - \nabla_j f(x^{k-d_k}) + M\sqrt{\rho_\tau \tau}(x_j^{k+1} - x_j^k) & \text{if } j = j_k; \\ \frac{1}{\gamma_j^k}(x_j^{l(j,k)} - x_j^{k+1}) + \nabla_j f(x^{k+1}) - \nabla_j f(x^{l(k,j)-d_{l(j,k)}}) & \text{otherwise;} \end{cases}$$

$$B^k = \begin{bmatrix} M\frac{\sqrt{\rho_\tau}(\tau-1)}{\sqrt{\tau}}(x^k - x^{k-1}) \\ \vdots \\ M\frac{\sqrt{\rho_\tau}}{\sqrt{\tau}}(x^{k-\tau+2} - x^{k-\tau+1}) \end{bmatrix},$$

then a quick look at optimality conditions verifies $(A_1^k, \ldots, A_m^k, B^k) \in \partial_L \Phi(z^{k+1})$ and, if we define $C^k := (A_1^k, \ldots, A_m^k) - M\sqrt{\rho_\tau \tau}(x^{k+1} - x^k)$, then $C^k \in \partial_L \Psi(x^{k+1})$. In addition, there exists a constant c_0 such that

$$\|(A_1^k, \ldots, A_m^k, B^k)\| \leq c_0 \sum_{h=k-\tau-K}^{k} \|x^{h+1} - x^h\| \to 0. \tag{5.2}$$

(In particular, $C^k \to 0$, too.) These limits also imply that all cluster points points are stationary points of Φ—provided that, for every converging subsequence $x^{k_q} \to x$, we have $\Phi(z^{k_q}) \to \Phi(x, \ldots, x)$.

To show this, we follow the same path as we did in the stochastic case: Fix a cluster point $x \in \mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$, say $x^{k_q} \to x$. Then

$$z^{k_q} \to (x, \ldots, x); \qquad z^{l(k_q,j)} \to (x, \ldots, x); \qquad \lim_{q\to\infty} f(x^{k_q}) = f(x).$$

Again, proving that $\lim_{q\to\infty} r_j(x_j^{k_q}) = r_j(x_j)$ is a little subtler because r_j is not continuous; it is merely lower semicontinuous.

For this, we use two properties: First, by lower semicontinuity, we have

$$\liminf_{q\to\infty} r_j(x_j^{k_q}) \geq r_j(x_j).$$

Second, by the definition of x_j^k as a proximal point, for all $y \in \mathcal{H}_j$ and $k \in \mathbb{N}$, we have

$$\begin{aligned} &r_j(x_j^k) + \langle \nabla_j f(x^{l(k,j)-d_{l(k,j)}}), x_j^k - x_j^{l(k,j)} \rangle + \frac{1}{2\gamma_j^{l(k,j)}} \|x_j^k - x_j^{l(k,j)}\|^2 \\ &\quad \leq r_j(y) + \langle \nabla_j f(x^{l(k,j)-d_{l(k,j)}}), y - x_j^{l(k,j)} \rangle + \frac{1}{2\gamma_j^{l(k,j)}} \|y - x_j^{l(k,j)}\|^2. \end{aligned}$$

In particular, by rearranging the above inequality for $k = k_q$ and $y = x_j$ and by taking a lim sup, we have

$$\begin{aligned} &\limsup_{q\to\infty} r_j(x_j^{k_q}) \\ &\leq \limsup_{q\to\infty} \left(r_j(x_j) + \langle \nabla_j f(x^{l(k_q,j)-d_{l(k_q,j)}}), x_j - x_j^k \rangle + \frac{1}{2\gamma_j^{l(k_q,j)}} \|x_j - x_j^{l(k_q,j)}\|^2 \right) \\ &\leq r_j(x_j). \end{aligned}$$

Therefore, $\lim_{q\to\infty} r_j(x_j^{k_q}) = r_j(x_j)$.

Altogether, because $\lim_{k\to\infty} \Phi(z^k) = \lim_{k\to\infty} f(x^k) + r(x^k)$, we have

$$\Phi(z^{k_q}) \to \Phi(x, \ldots, x) = f(x) + r(x) \qquad \text{and} \qquad \Psi(x^{k_q}) \to f(x) + r(x).$$

Moreover, the subgradients

$$(A_1^{k_q-1}, \ldots, A_m^{k_q-1}, B^{k_q-1}) \in \partial_L \Phi(z^{k_q}) \qquad \text{and} \qquad C^{k_q-1} \in \partial_L \Psi(x^{k_q})$$

converge to zero. Therefore, $0 \in \partial_L \Phi(x, \ldots, x)$ and $0 \in \partial_L \Psi(x)$.

We finish the proof of Part 1 with the proof of (5.1).

Proof (of (5.1)) We bound the smooth term first:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla_{j_k} f(x^k), x_{j_k}^{k+1} - x_{j_k}^k \rangle + \frac{L_{j_k}(x_{-j_k}^k)}{2} \|x_{j_k}^{k+1} - x_{j_k}^k\|^2.$$

Next we bound the nonsmooth term:

$$r_{j_k}(x_{j_k}^{k+1}) \leq r_{j_k}(x_{j_k}^k) - \langle \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle - \frac{1}{2\gamma_{j_k}^k} \|x_{j_k}^{k+1} - x_{j_k}^k\|^2.$$

Both terms together now:

$$f(x^{k+1}) + \sum_{j=1}^m r_j(x_j^{k+1}) \leq f(x^k) + \sum_{j=1}^m r_j(x_j^k) + \langle \nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle \\ - \frac{1}{2}\left(\frac{1}{\gamma_{j_k}^k} - L_{j_k}(x_{-j_k}^k)\right)\|x_{j_k}^{k+1} - x_{j_k}^k\|^2.$$

The cross term needs care. In particular, the following sequence of inequalities is true for any $C > 0$:

$$\begin{aligned}
&\langle \nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle \\
&\leq M\|x^k - x^{k-d_k}\|\|x_{j_k}^{k+1} - x_{j_k}^k\| \qquad \text{(by Assumption 1)} \\
&\leq \frac{M^2}{2C}\|x^k - x^{k-d_k}\|^2 + \frac{C}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \\
&\leq \frac{M^2\rho_\tau}{2C}\sum_{j=1}^m \sum_{h=k-d_{k,j}+1}^k \|x_j^h - x_j^{h-1}\|^2 + \frac{C}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \qquad \text{(by Jensen's inequality)} \\
&\leq \frac{M^2\rho_\tau}{2C}\sum_{j=1}^m \sum_{h=k-\tau+1}^k \|x_j^h - x_j^{h-1}\|^2 + \frac{C}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \\
&= \left(\frac{M^2\rho_\tau}{2C}\sum_{h=k-\tau+1}^k (h-k+\tau)\|x^h - x^{h-1}\|^2 - \frac{M^2\rho_\tau}{2C}\sum_{h=k-\tau+2}^{k+1}(h-(k+1)+\tau)\|x^h - x^{h-1}\|^2\right) \\
&\quad + \frac{M^2\rho_\tau\tau}{2C}\|x^{k+1} - x^k\|^2 + \frac{C}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2.
\end{aligned}$$

We collect all these alternating terms in the sequence $\{\kappa_k\}_{k\in\mathbb{N}}$, defined by

$$\kappa_k := \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}}\sum_{h=k-\tau+1}^k (h-k+\tau)\|x^h - x^{h-1}\|^2,$$

and set $C = M\sqrt{\rho_\tau\tau}$. Thus, because $\|x^{k+1} - x^k\|^2 = \|x_{j_k}^{k+1} - x_{j_k}^k\|^2$, we have

$$\begin{aligned}
\kappa_{k+1} &\leq \kappa_k - \langle \nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle + \frac{M^2\rho_\tau\tau}{2C}\|x^{k+1} - x^k\|^2 + \frac{C}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \\
&= \kappa_k - \langle \nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle + \left(\frac{M^2\rho_\tau\tau}{2C} + \frac{C}{2}\right)\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \\
&= \kappa_k - \langle \nabla_{j_k} f(x^k) - \nabla_{j_k} f(x^{k-d_k}), x_{j_k}^{k+1} - x_{j_k}^k \rangle + M\sqrt{\rho_\tau\tau}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&f(x^{k+1}) + \sum_{j=1}^m r_j(x_j^{k+1}) + \kappa_{k+1} \\
&\leq f(x^k) + \sum_{j=1}^m r_j(x_j^k) + \kappa_k - \frac{1}{2}\left(\frac{1}{\gamma_{j_k}^k} - L_{j_k}(x_{-j_k}^k) - 2M\sqrt{\rho_\tau\tau}\right)\|x_{j_k}^{k+1} - x_{j_k}^k\|^2.
\end{aligned}$$

In particular for all $k \in \mathbb{N}$, we have $\Phi(z^k) = f(x^k) + \sum_{j=1}^m r_j(x_j^k) + \kappa_k$, so (5.1) follows. □

Part 2: Let C denote the limit of $\Psi(x^k)$ and $\Phi(z^k)$ as $k \to \infty$ (which exists by Part 1), let $x \in \mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$, and suppose that $x^{k_q} \to x$. Then $C = \lim_{q\to\infty} \Psi(x^{k_q}) = \Psi(x) = \Phi(x, \ldots, x) = \lim_{q\to\infty} \Phi(z^{k_q})$. Thus, Φ (respectively Ψ) is constant on $\mathcal{C}(\{z^k\}_{k\in\mathbb{N}})$ (respectively $\mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$).

The bound on the limit of the objective value is a consequence of (5.1): First,

$$\Phi(x^0, x^{-1}, \ldots, x^{-\tau}) = \Psi(x^0).$$

Second, we have

$$\begin{aligned} (\forall k \in \mathbb{N}) \qquad & \Phi(x^{k+1}, x^k, \ldots, x^{k-\tau+1}) \le \Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) - Y_k; \\ \implies & \Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) \le \Psi(x^0) - \sum_{i=0}^{k-1} Y_i. \end{aligned} \tag{5.3}$$

Finally, we have only the strict decrease property left to prove: If x^0 is not a stationary point, then for some $k \le K$, we have $Y_k = \|x_{j_k}^{k+1} - x_{j_k}^k\| > 0$. Thus, the decrease property follows from (5.3) and the limit: $\lim_{k\to\infty} \Psi(x^k) = \lim_{k\to\infty} \Phi(z^k) < \Psi(x^0)$.

Finally, we return to the level set inclusion, which now follows easily from (5.1) (which does not depend on the boundedness of the iterates):

$$\Psi(x^k) \le \Phi(x^k, x^{k-1}, \ldots, x^{k-\tau}) \le \Psi(x^0).$$

□

Remark 5.1 The connectedness and compactness of $\mathcal{C}(\{x^k\}_{k\in\mathbb{N}})$ are implied by the limit $x^k - x^{k+1} \to 0$; see [5, Remark 3.3] for details.

Equation (5.1) figures again below, so we isolate the main content here:

Corollary 5.1 (A Decreasing Function Value Bound) *Regardless of whether $\{x^k\}_{k\in\mathbb{N}}$ is bounded, there exists $C > 0$ such that for all $k \in \mathbb{N}$, we have the following bound:*

$$\Phi(z^{k+1}) \le \Phi(z^k) - C\|x^{k+1} - x^k\|^2. \tag{5.4}$$

5.2 Global Sequence Convergence

The following Uniformized KL property is key to proving that $\{z^k\}_{k\in\mathbb{N}}$ converges.

Theorem 5.2 (Uniformized KL Property [5, Lemma 3.6]) *Let Q be a compact set, let $g : \mathcal{H} \to (-\infty, \infty]$ be proper, lower semicontinuous function that is constant on Q and satisfies the KL property at every point of Q. Then there exists $\varepsilon > 0, \eta > 0$, and $\varphi \in F_\eta$, such that for all $\overline{u} \in Q$ and all u in the intersection*

$$\{u \in \mathcal{H} \mid \operatorname{dist}(u, Q) < \varepsilon\} \cap \{u \in \mathcal{H} \mid g(\overline{u}) < g(u) < g(\overline{u}) + \eta\}, \tag{5.5}$$

we have

$$\varphi'(g(u) - g(\overline{u}))\operatorname{dist}(0, \partial_L g(u)) \ge 1.$$

With the uniformized KL property in hand, we can prove that $\{z^k\}_{k\in\mathbb{N}}$ has finite length and, hence, converges.

Theorem 5.3 (A Finite Length Property) *Suppose that $\{x^k\}_{k\in\mathbb{N}}$ is bounded and that Φ is a KL function. Then*

1. *The sequence $\{z^k\}_{k\in\mathbb{N}}$ has finite length, i.e.,*
$$\sum_{k=0}^{\infty} \|z^{k+1} - z^k\| < \infty.$$
2. *The sequence $\{z^k\}_{k\in\mathbb{N}}$ converges to a stationary point of Φ, and the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to a stationary point of Ψ.*

Proof **Part 1:** Let z be any cluster point of $\{z^k\}$. Then as we argued in Theorem 5.1, the following limit holds:
$$\lim_{k\to\infty} \Phi(z^k) = \Phi(z). \tag{5.6}$$

The sequence $\Phi(z^k)$ is decreasing, so if for some $\overline{k} \in \mathbb{N}$ we have $\Phi(z^{\overline{k}}) = \Phi(z)$, then $\Phi(z^k) = \Phi(z)$ for all $k \geq \overline{k}$. In that case, after applying (5.4) τ times, we find that there is a constant $C > 0$ such that for any $k \geq \overline{k}$, we have
$$C\|z^{k+\tau+1} - z^{k+\tau}\|^2 \leq \Phi(z^k) - \Phi(z^{k+\tau+1}) = 0$$
and moreover, by a simple induction, we find that $\{z^k\}_{k\in\mathbb{N}}$ must be eventually constant and, therefore, be of finite length.

On the other hand, if no such $\overline{k}$ exists (and every z^k is non-stationary), then for all $k \in \mathbb{N}$, we have $\Phi(z^k) > \Phi(z)$. Let $k_0 \in \mathbb{N}$ be large enough such that (for the ε and η in Theorem 5.2)
$$(\forall k \geq k_0) \qquad \Phi(z^k) < \Phi(z) + \eta \qquad \text{and} \qquad \operatorname{dist}\left(z^k, \mathcal{C}\left(\{z^k\}_{k\in\mathbb{N}}\right)\right) < \varepsilon \tag{5.7}$$
Then z^k belongs to the intersection in (5.5) with $Q = \mathcal{C}(\{z^k\}_{k\in\mathbb{N}})$ as soon as $k \geq k_0$, and Q is compact by Remark 5.1.

Now let $\varphi \in F_\eta$ be the concave continuous function from Theorem 5.2. Then, for $k \geq k_0$, we have
$$\varphi'(\Phi(z^k) - \Phi(z))\operatorname{dist}(\partial_L \Phi(z^k), 0) \geq 1.$$
Each of the terms in this product can be simplified. First, because φ is concave and by the bound in Corollary 5.4, we have
$$\begin{aligned}\varphi(\Phi(z^k) - \Phi(z)) - \varphi(\Phi(z^{k+1}) - \Phi(z)) &\geq \varphi'(\Phi(z^k) - \Phi(z))(\Phi(z^k) - \Phi(z^{k+1})) \\ &\geq \varphi'(\Phi(z^k) - \Phi(z))C\|x^{k+1} - x^k\|^2.\end{aligned}$$
Second, from (5.2), there exists $c_0 > 0$ such that
$$\varphi'(\Phi(z^k) - \Phi(z)) \geq \frac{1}{\operatorname{dist}(0, \partial_L \Phi(z^k))} \geq \frac{1}{c_0 \sum_{h=k-\tau-K-1}^{k-1} \|x^{h+1} - x^h\|}.$$

Altogether, with

$$(\forall k \geq k_0) \qquad \epsilon_{k-k_0} := \frac{C}{c_0}\left(\varphi(\Phi(z^k) - \Phi(z)) - \varphi(\Phi(z^{k+1}) - \Phi(z))\right),$$

we have

$$(\forall k \geq k_0) \qquad \epsilon_{k-k_0} \geq \frac{\|x^{k+1} - x^k\|^2}{\sum_{h=k-\tau-K-1}^{k-1} \|x^{h+1} - x^h\|},$$

and, moreover, $\sum_{k=0}^{\infty} \epsilon_k < \infty$. Rearranging, we find that

$$\begin{aligned}\|x^{k+1} - x^k\| &\leq \sqrt{\left(\sum_{h=k-\tau-K-1}^{k-1} \|x^{h+1} - x^h\|\right)\epsilon_{k-k_0}} \\ &\leq \frac{1}{2(\tau+K+1)}\left(\sum_{h=k-\tau-K-1}^{k-1} \|x^{h+1} - x^h\|\right) + \frac{(\tau+K+1)}{2}\epsilon_{k-k_0}.\end{aligned}$$

Thus, to show that the sequence has finite length we apply the following Lemma with $a_{k-k_0} = \|x^k - x^{k-1}\|$ and $b_i \equiv (2(\tau+K+1))^{-1}$, which shows that $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < \infty$ and, consequently, that $\sum_{k=0}^{\infty} \|z^{k+1} - z^k\| < \infty$.

Lemma 5.1 *Let $\{\epsilon_k\}_{k\in\mathbb{N}}$ be a summable sequence, let $b_0, \ldots, b_{\tau+K}$ be a sequence of nonegative real numbers such that $\sum_{i=0}^{\tau+K} b_i < 1$, and let $\{a_k\}_{k\in\mathbb{N}}$ be a sequence of nonnegative real numbers (extended to $\mathbb{Z}$ by $a_{-k} := a_0$ for all $k \in \mathbb{N}$) such that for all $k \in \mathbb{N}$, we have $a_{k+1} \leq \sum_{h=k-\tau-K-1}^{k-1} b_{k+\tau+K+1-h} a_{h+1} + \epsilon_k$. Then $\sum_{k=0}^{\infty} a_k < \infty$.*

This Lemma is a straightforward generalization of [6, Lemma 3], so we omit its proof.

Part 2: Sequences of finite length are known to be Cauchy and, hence, convergent. Therefore, the sequence $\{z^k\}_{k\in\mathbb{N}}$ converges. By Theorem 5.1 the limit of $\{z^k\}_{k\in\mathbb{N}}$ limit is a stationary point of Φ, while the limit of $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of Ψ. □

5.3 Convergence rates

For convergence rate analysis, the class of semi-algebraic functions (Definition 3.1), which are known to be KL functions, are the easiest to get a handle on. It turns out that Algorithm 2 can converge in a finite number of steps, linearly, or sublinearly, depending on a certain exponent θ defined below, whenever Ψ is semi-algebraic.

Theorem 5.4 (Convergence Rates) *Suppose that $\{x^k\}_{k\in\mathbb{N}}$ is bounded and that Φ is a KL function. Let $z = (x, \ldots, x) \in \mathcal{H}^{1+\tau}$ be the limit of $\{z^k\}_{k\in\mathbb{N}}$ (which exists by Theorem 5.3). Then*

1. *In general,*

$$\min_{t=0,\ldots,k} \operatorname{dist}(0, \partial_L \Phi(z^t)) = o\left(\frac{1}{k+1}\right) \qquad \text{and} \qquad \min_{t=0,\ldots,k} \operatorname{dist}(0, \partial_L \Psi(z^t)) = o\left(\frac{1}{k+1}\right).$$

2. *Suppose Ψ is semi-algebraic. Then Φ is semi-algebraic, it satisfies the KL inequality with $\varphi(s) := cs^{(1-\theta)}$, where $\theta \in [0,1)$ and $c > 0$, and*
   (a) if $\theta = 0$, then we have $0 \in \partial_L \Phi(z^k)$ and $0 \in \partial_L \Psi(x^k)$ for all sufficiently large $k \in \mathbb{N}$;
   (b) if $\theta \in (0, 2^{-1}]$, then there exists $\rho \in (0,1)$ such that

$$\Psi(x^k) - \Psi(x) \leq \Phi(z^k) - \Phi(z) = O\left(\rho^{\lfloor \frac{k-k_1}{(\tau+K+1)} \rfloor}\right);$$

   (c) if $\theta \in (2^{-1}, 1)$, then

$$\Psi(x^k) - \Psi(x) \leq \Phi(z^k) - \Phi(z) = O\left(\frac{1}{(k+1)^{\frac{1}{2\theta-1}}}\right).$$

Proof **Part 1:** The finite length property of z^k, shown in Theorem 5.3, implies that $\min_{t=0,\ldots,k} \|z^t - z^{t+1}\| = o((k+1)^{-1})$; see [11, Part 4 of Lemma 3]. Therefore, from (5.2), we have

$$\operatorname{dist}(0, \partial_L \Phi(z^k)) \leq \|(A_1^{k-1}, \ldots, A_m^{k-1}, B^{k-1})\| = o\left(\frac{1}{k+1}\right);$$
$$\text{and} \qquad \operatorname{dist}(0, \partial_L \Psi(x^k)) \leq \|C^{k-1}\| = o\left(\frac{1}{k+1}\right).$$

Part 2: The class of semi-algebraic functions is closed under addition. Therefore, because Ψ is semi-algebraic and $\Phi - \Psi$ is semi-algebraic (when Ψ is viewed as a function on $\mathcal{H}^{1+\tau}$ in the obvious way), it follows that Φ is semi-algebraic. The claimed form of φ follows from [2, Section 4.3].

Now we assume that $\{z^k\}_{k\in\mathbb{N}}$ does not converge in finitely many steps; if it did converge in only finitely many steps, all the claimed results clearly hold. As in the proof of Theorem 5.3, we choose k_0 large enough that (5.7) holds, and we consider only $k \geq k_0$.

We use the shorthand $\Phi_k = \Phi(z^k) - \Phi(z)$, where z is the unique limit point of $\{z^k\}_{k\in\mathbb{N}}$. Then, by Corollary 5.1, we have

$$\Phi_k - \Phi_{k+\tau+K+1} \geq C\left(\sum_{h=k}^{k+\tau+K} \|x^{h+1} - x^h\|^2\right) \geq \frac{C}{\tau+K+1}\left(\sum_{h=k}^{k+\tau+K} \|x^{h+1} - x^h\|\right)^2.$$

In addition, as in the proof of (5.3), we have

$$c(1-\theta)\Phi_{k+\tau+K+1}^{-\theta} = \varphi'(\Phi_{k+\tau+K+1}) \geq \frac{1}{\operatorname{dist}(0, \partial_L \Phi(z^{k+\tau+K+1}))} \geq \frac{1}{c_0 \sum_{h=k}^{k+\tau+K} \|x^{h+1} - x^h\|}. \tag{5.8}$$

Therefore, we have

$$(\forall k \geq k_0) \qquad \Phi_k - \Phi_{k+\tau+K+1} \geq C_1 \Phi_{k+\tau+K+1}^{2\theta}. \tag{5.9}$$

where $C_1 := C(c^2(1-\theta)^2 c_0^2 (K+\tau+1))^{-1}$.

Part 2a: Suppose that $\theta = 0$. Then for all $k \geq k_0$, we have $\Phi_k - \Phi_{k+\tau+K+1} \geq C_1 > 0$, which cannot hold because $\Phi_k \to 0$. Thus, $\{\Phi(z^k)\}_{k\in\mathbb{N}}$ must converge in finitely many steps, and, by the first inequality of the proof of Theorem 5.3, this implies that $\{z^k\}_{k\in\mathbb{N}}$ converges to a stationary point of Φ in finitely many steps. (In particular, x^k also converges to a stationary point of Ψ, by Part 1 of Theorem 5.1.)

Part 2b: Suppose that $\theta \in (0, 2^{-1}]$. Choose $k_1 \geq k_0$ large enough that $\Phi_k^{2\theta} \geq \Phi_k$ (such a k_1 exists because $\Phi_k \to 0$). Then

$$(\forall k \geq k_1 + \tau + K + 1) \qquad \Phi_k \leq \frac{1}{1+C_1}\Phi_{k-K-\tau-1} \implies \Phi_k \leq \left(\frac{1}{1+C_1}\right)^{\lfloor \frac{k-k_1}{(\tau+K+1)} \rfloor} \Phi_{k_1},$$

where we use that Φ_k is nonincreasing.

Part 2c: Suppose that $\theta \in (2^{-1}, 1)$. Let $h : (0,\infty) \to (0,\infty)$ be the nonincreasing function $h(s) := s^{-2\theta}$. Then from (5.9) we find that

$$\begin{aligned} C_1 \leq \frac{h(\Phi_{k+\tau+K+1})}{h(\Phi_k)}(\Phi_k - \Phi_{k+\tau+K+1})h(\Phi_k) &\leq \frac{h(\Phi_{k+\tau+K+1})}{h(\Phi_k)} \int_{\Phi_{k+\tau+K+1}}^{\Phi_k} h(s)ds \\ &= \frac{h(\Phi_{k+\tau+K+1})}{h(\Phi_k)} \frac{\Phi_{k+\tau+K+1}^{1-2\theta} - \Phi_k^{1-2\theta}}{2\theta - 1}. \end{aligned}$$

Let $R \in (1,\infty)$ be a fixed number. We will deal with the troublesome ratio $h(\Phi_{k+\tau+K+1})(h(\Phi_k))^{-1}$ with two cases.

Case 1: $h(\Phi_{k+\tau+K+1})(h(\Phi_k))^{-1} \leq R$. In this case

$$\frac{C_1}{R} \leq \frac{\Phi_{k+\tau+K+1}^{1-2\theta} - \Phi_k^{1-2\theta}}{2\theta - 1}.$$

Case 2: $h(\Phi_{k+\tau+K+1})(h(\Phi_k))^{-1} > R$. In this case, we set $q := R^{-1/2\theta} \in (0,1)$ and deduce the bounds

$$\Phi_{k+\tau+K+1}^{1-2\theta} > q^{1-2\theta}\Phi_k^{1-2\theta} \implies (q^{1-2\theta} - 1)\Phi_k^{1-2\theta} \leq \Phi_{k+\tau+K+1}^{1-2\theta} - \Phi_k^{1-2\theta}.$$

Choose $k_1 \in \mathbb{N}$ such that $k_1 \geq k_0$ and $(q^{1-2\theta} - 1)\Phi_k^{1-2\theta} > C_1 R^{-1}$ (such a k_1 exists because $\Phi_k \to 0$).

Thus, we have the following bounds for all $t \in \mathbb{N}$:

$$\begin{aligned} (\forall k \geq k_1) \qquad \frac{C_1}{R} &\leq \frac{\Phi_{k+\tau+K+1}^{1-2\theta} - \Phi_k^{1-2\theta}}{2\theta - 1} \\ \implies t\frac{C_1}{R} &\leq \frac{\Phi_{k+t(\tau+K+1)}^{1-2\theta} - \Phi_k^{1-2\theta}}{2\theta - 1} \\ \implies \Phi_{k+t(\tau+K+1)} &\leq \left(\frac{1}{C_1 t(2\theta-1)R^{-1} + \Phi_k^{1-2\theta}}\right)^{\frac{1}{2\theta-1}} \\ &\leq \left(\frac{1}{C_1 t(2\theta-1)R^{-1} + \Phi_{k_1}^{1-2\theta}}\right)^{\frac{1}{2\theta-1}}, \end{aligned}$$

which implies the claimed bound:

$$(\forall k \geq k_1) \qquad \Phi_k \leq \left(\frac{1}{C_1 \lfloor \frac{k-k_1}{\tau+K+1} \rfloor (2\theta-1)R^{-1} + \Phi_{k_1}^{1-2\theta}}\right)^{\frac{1}{2\theta-1}} = O\left(\frac{1}{(k+1)^{\frac{1}{2\theta-1}}}\right). \qquad \square$$

6 Discussion

In this section, we lay out assumptions under which Asynchronous PALM converges. It is likely that weaker assumptions suffice for your favorite model, but let us see how far we can get with the stricter assumptions that we propose—if only to make it easier to design software capable of solving (1.1) for several problems all at once.

Ensuring Boundedness with Coercivity. To get anywhere in our results, we must assume that the $\{x^k\}_{k\in\mathbb{N}}$ is bounded. In both the stochastic and deterministic cases there is a sequence $\{z^k\}_{k\in\mathbb{N}}$ that is bounded if, and only if, $\{x^k\}_{k\in\mathbb{N}}$ is bounded, and a Lyapunov function Φ, which, for all $k \in \mathbb{N}$, satisfies one of the following inequalities

$$\begin{aligned}(4.1) &\implies \mathbb{E}\left[\Phi(z^{k+1}) \mid \mathcal{F}_k\right] \leq \Phi(z^k)\\ (5.4) &\implies \Phi(z^{k+1}) \leq \Phi(z^k),\end{aligned}$$

regardless of whether $\{z^k\}_{k\in\mathbb{N}}$ is bounded. If the expectation bound holds, the supermartingale convergence theorem (quoted in Theorem 4.2) implies that the term $\{\Phi(z^{k+1})\}_{k\in\mathbb{N}}$ is almost surely bounded; similarly, if the deterministic inequality holds, then $\{\Phi(z^k)\}_{k\in\mathbb{N}}$ is bounded. Thus, we turn our attention to conditions under which the boundedness of $\{\Phi(z^k)\}_{k\in\mathbb{N}}$ implies the boundedness of $\{z^k\}_{k\in\mathbb{N}}$ (we now ignore the distinction between almost sure boundedness and deterministic boundedness).

In such a general context, the easiest condition to verify is *coercivity* of Ψ:

$$\lim_{\|z\|\to\infty} \Psi(z) = \infty.$$

If coercivity holds, then clearly the boundedness of $\{\Phi(z^k)\}_{k\in\mathbb{N}}$ and the bound $\Psi(x^k) \leq \Phi(z^k)$ implies the boundedness of $\{x^k\}_{k\in\mathbb{N}}$ and $\{z^k\}_{k\in\mathbb{N}}$. Thus, to ensure boundedness of $\{x^k\}_{k\in\mathbb{N}}$, the most general assumption we employ is that Ψ is coercive.

Ensuring the KL Property with Semi-Algebraicity. To prove that the Lyapunov function Φ has the KL property, it is not necessarily enough to show that Ψ has the KL property. However, because the class of *semi-algebraic* functions (see Definition 3.1) is closed under addition and $\Phi - \Psi$ is semi-algebraic, it follows that

$$\Psi \text{ semi-algebraic} \implies \Phi \text{ semi-algebraic}.$$

Thus, to ensure Φ is a KL function, the most general assumption we employ is that Ψ is semi-algebraic.

Ensuring Bounded Lipschitz Constants. We must assume that $L_j(x^k_{-j})$ is bounded for all k and j and that ∇f has Lipschitz constant M on the set of iterates $\{x^k\}_{k\in\mathbb{N}} \cup \{x^{k-d_k}\}_{k\in\mathbb{N}}$. This set is not necessarily bounded, but when Ψ is coercive, we can choose M to be the Lipschitz constant of ∇f on the minimal box B containing $\{x \mid \Psi(x) \leq \Psi(x^0)\}$, and with that choice of M, one can check by induction that x^k will indeed stay in B. In the stochastic case, we cannot guarantee that the iterates lie in the level set, so a similar argument is unavailable.

Using Linesearch. A quick look verifies that all results of Section 5 continue to hold as long as we choose γ_j^k in such a way that there exists $C > 0$ with the property that for all $k \in \mathbb{N}$, we have

$$\Phi(z^{k+1}) \leq \Phi(z^k) - C\|x^{k+1} - x^k\|^2.$$

Thus, the following is a valid line search criteria: given x^k, choose $\gamma > 0$ so that for

$$x_j^{k+1} \in \begin{cases} \mathbf{prox}_{\gamma r_j}(x_j^k - \gamma \nabla_j f(x^{k-d_k})) & \text{if } j = j_k; \\ \{x_j^k\} & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} f(x^{k+1}) + r(x^{k+1}) + \left(C + \frac{M\sqrt{\rho_\tau \tau}}{2}\right)\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 \\ \leq f(x^k) + r(x^k) + \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}} \sum_{h=k-\tau+1}^{k} \|x^h - x^{h-1}\|^2. \end{aligned}$$

Importantly, we can quickly update the sum $\xi_k := \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}} \sum_{h=k-\tau+1}^{k} \|x^h - x^{h-1}\|^2$ by storing the τ numbers $\frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}}\|x^k - x^{k-1}\|^2, \ldots, \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}}\|x^{k-\tau+1} - x^{k-\tau}\|^2$:

$$\xi_{k+1} = \xi_k + \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 - \frac{M\sqrt{\rho_\tau}}{2\sqrt{\tau}}\|x^{k-\tau+1} - x^{k-\tau}\|^2.$$

Thus, with coercivity and the KL property in hand, we have the following theorem:

Theorem 6.1 (Global Convergence of Deterministic Asynchronous PALM) *Suppose that Ψ is coercive, semi-algebraic, and ∇f is M-Lipschitz continuous on the minimal box B containing the level set $\{x \mid \Psi(x) \leq \Psi(x^0)\}$. Then $\{x^k\}_{k\in\mathbb{N}}$ from Algorithm 2 globally converges to a stationary point of Ψ.*

6.1 Example: Generalized Low Rank Models

A broad family of models with which hidden low rank structure of data may be discovered, analyzed, and sometimes, enforced, has been outlined in the Generalized Low Rank Model (GLRM) framework proposed in [27]. The original PALM [5] algorithm was motivated by the most fundamental of all GLRMs, namely matrix factorization, and since the time that PALM was introduced, the authors of [27] have used this approach quite successfully to optimize other, more general low rank models.

Model. In a GLRM, you are given a mixed data type matrix $A \in T^{d_1 \times d_2}$, which has, for example, real, boolean, or categorical entries represented by T. In the theology of GLRMs, we imagine that there are two collections of vectors $\{x_{i,1}\}_{i=0}^{d_1} \subseteq \mathbb{R}^d$ and $\{x_{l,2}\}_{l=0}^{d_2} \subseteq \mathbb{R}^d$ for which, in the case of a real-valued matrix A, we have $\langle x_{i,2}, x_{l,2}\rangle \approx A_{il}$; but in general there is a differentiable loss function $f_{il}(\cdot, A_{il}) : \mathbb{R} \to \mathbb{R}$, with L_{il}-Lipschitz continuous derivative f'_{il}, that measures how closely $\langle x_{i,2}, x_{l,2}\rangle$ represents A_{il}. Then we define the global loss function from these local terms:

$$f(x_{1,1}, \ldots, x_{d_1,1}, x_{1,2}, \ldots, x_{d_2,2}) := \sum_{i=1}^{d_1} \sum_{l=1}^{d_2} f_{il}(\langle x_{i,1}, x_{l,2}\rangle; A_{il}).$$

For the special case of real-valued matrix factorization, the local terms are all identical and equal to $f_{il}(a, A_{il}) := 2^{-1}(a - A_{il})^2$ and the Lipschitz constant is $L_{il} \equiv 1$.

GLRMs gain a lot of biasing power from adding nonsmooth, nonconvex regularizers $r_{i,1}, r_{l,2} : \mathbb{R}^d \to \mathbb{R}$ to the global loss function f which, after renaming $x := (x_{1,1}, \ldots, x_{d_1,1}, x_{1,2}, \ldots, x_{d_2,2}) \in \mathbb{R}^{d \times (d_1 + d_2)}$, leads to the final objective function:

$$\Psi(x) := f(x) + \sum_{i=1}^{d_1} r_{i,1}(x_{i,1}) + \sum_{l=1}^{d_2} r_{l,2}(x_{l,2}).$$

Lipschitz Constants. The component-wise Lipschitz constants of the partial gradients (we just look at the $x_{i,1}$ components; the other case is symmetric)

$$\nabla_{x_{i,1}} f(x) = \sum_{l=1}^{d_2} x_{l,2} f'_{il}(\langle x_{i,1}, x_{l,2}\rangle; A_{il})$$

are easily seen to be $\sum_{l=1}^{d_2} \|x_{l,2}\| L_{il}$. Thus, if $\{x^k\}_{k \in \mathbb{N}}$ is a bounded sequence, then the Lipschitz constants $L_j(x^k_{-j})$ remain bounded for all j and k. Further, simple probing reveals that ∇f is Lipschitz on bounded sets.

Coercivity. Among the special cases of GLRM objectives, coercivity holds, for example, for all variants of PCA, all variants of matrix factorization, quadratic clustering and mixtures, and subspace clustering.

KL Property via Semi-Algebraicity. Among the special cases of GLRM objectives, semi-algebraicity holds, for example, for standard, quadratically regularized, and sparse PCA; nonnegative, nonnegative orthogonal, and max norm matrix factorization; quadratic clustering; quadratic mixtures; and subspace clustering.

Thus, if they are semi-algebraic and cocoercive, GLRMs form a perfect set of examples for the PALM algorithm, and more generally, our Asynchronous PALM algorithm. Likely, most of the GLRMs considered in [27] will also meet the general KL assumption (as opposed to semi-algebraicity), however, verifying this condition requires a bit more work, in a direction orthogonal to the direction of this paper.

7 Conclusion

The Asynchronous PALM algorithm minimizes our model problem (1.1) by allowing asynchronous parallel inconsistent reading of data—an algorithmic feature that, when implemented on an n core computer, often speeds up algorithms by a factor proportional to n.

Problem (1.1) is a relatively simple nonsmooth, nonconvex optimization problem, but it figures prominently in the GLRM framework. A yet to be realized extension of this work might complicate our model problem (1.1) by letting each of the regularizers r_j depend on more than one of the optimization variables. Such an extension would significantly extend the reach of first-order algorithms in nonsmooth, nonconvex optimizations.

Acknowledgements: We thank Brent Edmunds, Robert Hannah, and Professors Madeleine Udell and Stephen J. Wright for helpful comments.

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming **116**(1), 5–16 (2007). DOI 10.1007/s10107-007-0133-5. URL http://dx.doi.org/10.1007/s10107-007-0133-5
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz inequality. Mathematics of Operations Research **35**(2), 438–457 (2010)
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 1st edn. Springer Publishing Company, Incorporated (2011)
4. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, vol. 23
5. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146**(1-2), 459–494 (2014)
6. Boţ, R.I., Csetnek, E.R.: An Inertial Tseng's Type Proximal Algorithm for Nonsmooth and Nonconvex Optimization Problems. Journal of Optimization Theory and Applications pp. 1–17 (2015)
7. Boţ, R.I., Csetnek, E.R.: Proximal-gradient algorithms for fractional programming. arXiv preprint arXiv:1601.08166 (2016)
8. Boţ, R.I., Csetnek, E.R., László, S.C.: An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. EURO Journal on Computational Optimization pp. 1–23 (2014)
9. Chouzenoux, E., Pesquet, J.C., Repetti, A.: A block coordinate variable metric forward-backward algorithm (2013)
10. Davis, D.: SMART: The Stochastic Monotone Aggregated Root-Finding Algorithm. arXiv preprint arXiv:1601.00698 (2016)
11. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. In: R. Glowinski, S. Osher, W. Yin (eds.) Splitting Methods in Communication and Imaging, Science and Engineering, p. Chapter 4. Springer (2016)
12. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting Methods with Variable Metric for Kurdyka–Łojasiewicz Functions and General Convergence Rates. Journal of Optimization Theory and Applications **165**(3), 874–900 (2015)
13. Hesse, R., Luke, D.R., Sabach, S., Tam, M.K.: Proximal Heterogeneous Block Implicit-Explicit Method and Application to Blind Ptychographic Diffraction Imaging. SIAM Journal on Imaging Sciences **8**(1), 426–457 (2015)
14. Li, G., Pong, T.K.: Douglas–rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Mathematical Programming pp. 1–31
15. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM Journal on Optimization **25**(4), 2434–2460 (2015)
16. Lian, X., Huang, Y., Li, Y., Liu, J.: Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization. In: Advances in Neural Information Processing Systems, pp. 2719–2727 (2015)
17. Liu, J., Wright, S.J., Ré, C., Bittorf, V., Sridhar, S.: An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. Journal of Machine Learning Research **16**, 285–322 (2015)
18. Liu, J., Wright, S.J., Sridhar, S.: An Asynchronous Parallel Randomized Kaczmarz Algorithm. arXiv preprint arXiv:1401.4780 (2014)
19. Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., Jordan, M.I.: Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. arXiv preprint arXiv:1507.06970 (2015)

20. Nesterov, Y.: Introductory Lectures on Convex Optimization : A Basic Course. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London (2004). URL http://opac.inria.fr/record=b1104789
21. Peng, Z., Wu, T., Xu, Y., Yan, M., Yin, W.: Coordinate Friendly Structures, Algorithms and Applications. arXiv preprint arXiv:1601.00863 (2016)
22. Peng, Z., Xu, Y., Yan, M., Yin, W.: ARock: an Algorithmic Framework for Asynchronous Parallel Coordinate Updates. arXiv preprint arXiv:1506.02396 (2015)
23. Recht, B., Re, C., Wright, S., Niu, F.: Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In: Advances in Neural Information Processing Systems, pp. 693–701 (2011)
24. Robbins, H., Siegmund, D.: A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications. In: Herbert Robbins Selected Papers, pp. 111–135. Springer (1985)
25. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer Science & Business Media (2009)
26. Tseng, P.: On the Rate of Convergence of a Partially Asynchronous Gradient Projection Algorithm. SIAM Journal on Optimization **1**(4), 603–619 (1991). DOI 10.1137/0801036. URL http://dx.doi.org/10.1137/0801036
27. Udell, M., Horn, C., Zadeh, R., Boyd, S.: Generalized Low Rank Models. arXiv preprint arXiv:1410.0342 (2014)
28. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. arXiv preprint arXiv:1410.1386 (2014)