Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems

Jérôme Bolte* Shoham Sabach† Marc Teboulle‡

Abstract

We introduce a proximal alternating linearized minimization (PALM) algorithm for solving a broad class of nonconvex and nonsmooth minimization problems. Building on the powerful Kurdyka-Łojasiewicz property, we derive a self-contained convergence analysis framework and establish that each bounded sequence generated by PALM globally converges to a critical point. Our approach allows to analyze various classes of nonconvex-nonsmooth problems and related nonconvex proximal forward-backward algorithms with semi-algebraic problem's data, the later property being shared by many functions arising in wide variety of fundamental applications. A by-product of our framework also shows that our results are new even in the convex setting. As an illustration of the results, we derive a new and simple globally convergent algorithm for solving the sparse nonnegative matrix factorization problem.

Key words: Alternating minimization, block coordinate descent, Gauss-Seidel method, Kurdyka-Łojasiewicz property, nonconvex-nonsmooth minimization, proximal forward-backward, sparse nonnegative matrix factorization.

Mathematics Subject Classification (2010): 90C05, 90C25, 90C30, 90C52, 65K05, 65F22, 49M37, 47J25.

1 Introduction

Minimizing the sum of a finite collections of given functions has been at the heart of mathematical optimization research. Indeed, such an abstract model is a convenient vehicle

*TSE (GREMAQ, Université Toulouse I), Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, France. E-mail: jerome.bolte@tse-fr.eu. This research benefited from the support of the FMJH Program Gaspard Monge in optimization and operation research (and from the support to this program from EDF) and it was co-funded by the European Union under the 7th Framework Programme "FP7-PEOPLE-2010-ITN", grant agreement number 264735-SADCO.

†School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: ssabach@post.tau.ac.il. This research was supported by a Tel Aviv University postdoctoral fellowship.

‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il. This research was partially supported by the Israel Science Foundation, ISF Grant 998-12.

which includes most practical models arising in a wide range of applications, whereby each function can be used to describe a specific required property of the problem at hand, either as an objective or as a constraint or both. Such a structure, while very general, still often allows one to beneficially exploit mathematical properties of the specific functions involved to devise simple and efficient algorithms. Needless to say that the literature in optimization research and its applications covering such a model is huge, and the present paper is not intended to review it. For some pioneering and early works that realized the potential of the sum optimization model, see for instance, Auslender [4], and Bertsekas and Tsitsiklis [12], with references therein.

Recently there has been a revived interest in the design and analysis of algorithms for solving optimization problems involving sum of functions, in particular is signal/image processing and machine learning. The main trend is solving very large scale problems, exploiting special structures/properties of the problem data toward the design of very simple scheme (*e.g.*, matrix/vector multiplications), yet capable of producing reasonable approximate solutions efficiently. In order to achieve these goals, the focus of this recent research has been with a particular emphasis on the development and analysis of algorithms for *convex models* which either describe a particular application at hand or is used as a relaxation for tackling an original nonconvex model. We refer the reader to the two very recent edited volumes [30] and [34] for a wealth of relevant and interesting works covering a broad spectrum of theory and applications which reflects this intense research activity.

In this work, we completely depart from the convex setting. Indeed, in many of the alluded applications, the original optimization model is often genuinely nonconvex and nonsmooth. This can be seen in a wide array of problems such as: compressed sensing, matrix factorization, dictionary learning, sparse approximations of signals and images, and blind decomposition, to mention just a few. We thus consider a broad class of nonconvex-nonsmooth problems of the form

$$(M) \qquad \text{minimize}_{x,y} \Psi(x,y) := f(x) + g(y) + H(x,y)$$

where the functions f and g are extended valued (*i.e.*, allowing the inclusion of constraints) and H is a smooth function (see more precise definitions in the next section). We stress that throughout this paper, no convexity whatsoever will be assumed in the objective or/and the constraints. Moreover, we note that the choice of two block of variables is for the sake of simplicity of exposition. Indeed, all the results derived in this paper hold true for a finite number of block-variables, see Section 3.6.

This model is rich enough to cover many of the applications mentioned above, and was recently studied in the work of Attouch *et al.* [2] which also provides the motivation of the present work. The standard approach to solve Problem (M) is via the so-called Gauss-Seidel iteration scheme, popularized in modern era under the name alternating minimization. That is, starting with some given initial point (x^0, y^0), we generate a sequence $\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$

via the scheme

$$x^{k+1} \in \operatorname{argmin}_x \Psi\left(x, y^k\right)$$
$$y^{k+1} \in \operatorname{argmin}_y \Psi\left(x^{k+1}, y\right).$$

Convergence results for the Gauss-Seidel method, also known as coordinate descent method, can be found in several studies, see *e.g.*, [4, 12, 29, 35]. One of the key assumptions necessary to prove convergence is that the minimum in each step is uniquely attained, see *e.g.*, [36]. Otherwise, as shown in Powell [32], the method may cycle indefinitely without converging. In the convex setting, for a continuously differentiable function Ψ, assuming strict convexity of one argument while the other is fixed, every limit point of the sequence $\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ generated by this method minimizes Ψ, see *e.g.*, [12]. Very recently, in [10], global rate of convergence results have been derived for block coordinate gradient projection algorithm for convex and smooth constrained minimization problems.

Removing the strict convexity assumption can be achieved by coupling the method with a proximal term, that is to consider the proximal regularization of the Gauss-Seidel scheme:

$$x^{k+1} \in \operatorname{argmin}_x \left\{\Psi\left(x, y^k\right) + \frac{c_k}{2}\left\|x - x^k\right\|^2\right\} \tag{1.1}$$
$$y^{k+1} \in \operatorname{argmin}_y \left\{\Psi\left(x^{k+1}, y\right) + \frac{d_k}{2}\left\|y - y^k\right\|^2\right\}, \tag{1.2}$$

where c_k and d_k are positive real numbers. In fact, such an idea was already suggested by Auslender in [6]. It was further studied in [7] with a nonquadratic proximal term to handle linearly constrained convex problems, and further results can be found in [19]. In all these works, only convergence of the subsequences can be established. In the nonconvex and nonsmooth setting, which is the focus of this paper, the situation becomes much harder, see *e.g.*, [35].

The present work is motivated by two very recent papers by Attouch *et al.* [2, 3], which appear to be the first works in the general nonconvex and nonsmooth setting, establishing in [2] convergence of the sequences generated by the proximal Gauss-Seidel scheme (see (1.1) and (1.2)), while in [3], a similar result was proven for the well-known proximal-forward-backward (PFB) algorithm applied to the nonconvex and nonsmooth minimization of the sum of a nonsmooth function with a smooth one (*i.e.*, Problem (M) with no y). Their approach relies on assuming that the objective function Ψ to be minimized satisfies the so-called Kurdyka-Łojasiewicz (KL) property [22, 25], which was developed for nonsmooth functions by Bolte *et al.* [16, 17] (see Section 2.4).

In both of these works, the suggested approach gains its strength from the fact that the class of functions satisfying the KL property is considerably large, and cover a wealth of nonconvex-nonsmooth functions arising in many fundamental applications, see more in the forthcoming Section 3 and in the Appendix.

Clearly, the scheme (1.1) and (1.2) always produce a nonincreasing sequence of function values, *i.e.*, for all $k \geq 0$ we have

$$\Psi\left(x^{k+1}, y^{k+1}\right) \leq \Psi\left(x^k, y^k\right)$$

and the sequence $\left\{\Psi\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ is bounded from below by $\inf\Psi$. Thus, with $\inf\Psi > -\infty$, the sequence $\left\{\Psi\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ converges to some real number, and as proven in [2], assuming that the objective function Ψ satisfies the KL property, every bounded sequence generated by the proximal regularized Gauss-Seidel scheme (1.1) and (1.2) converges to a critical point of Ψ. These are nice properties for the alluded scheme above. However, this scheme is *conceptual*, and not really a "true" algorithm, in the sense that it suffers from (at least) two main drawbacks. First, each step requires exact minimization of a nonconvex and nonsmooth problem. Secondly, it is a *nested* scheme which implies two nontrivial issues: (i) accumulations of computational errors in each step, and (ii) how and when to stop each step before passing to the next.

The above drawbacks motivates a very simple and naive approach, which can be traced back to [5] for smooth unconstrained minimization. Thus, for the more general Problem (M), for each block of coordinate perform one gradient step on the smooth part, while a proximal step is taken on the nonsmooth part. This idea contrasts with the entirely implicit step required by the proximal version of the Gauss-Seidel method (1.1) and (1.2), that is here, we consider an *approximation* of this scheme via the well-known and standard *proximal linearization* of each subproblem. This yields the Proximal Alternating Linearized Minimization (PALM) algorithm, whose exact description is given in Section 3.1. Thus, the root of our method can be viewed as nothing else but *an alternating minimization approach* for the so-called Proximal Forward-Backward (PFB) algorithm. Let us mention that the PFB algorithm has been extensively studied and successfully applied in many contexts in the convex setting, see *e.g.*, the recent monograph of Bauschke-Combettes [8] for a wealth of fundamental results and references therein.

Now, we briefly streamline the novelty of our approach and our contributions. First, the coupling of the Gauss-Seidel proximal scheme with PFB does not seem to have been analyzed in the literature within such a general nonconvex and nonsmooth setting proposed here. It allows to eliminate the difficulties evoked above with the scheme (1.1) and (1.2) and leads to a simple and tractable algorithm PALM, with global convergence results for nonconvex and nonsmooth semi-algebraic problems.

Secondly, while a part of the convergence result we develop in this article falls in the scope of a general convergence mechanism introduced and described in [3], we present here a self-contained thorough proof that avoids the use of these abstract results. The motivation stems from the fact that we target applications for which KL property holds at each point of the underlying space. Functions having this property are called KL functions. A very wide class of KL functions is provided by tame functions; these include in particular nonsmooth semi-algebraic and real subanalytic functions (see, *e.g.*, [2] and references therein). This property allows, through a "uniformization" result inspired by [1] (see Lemma 3.6) to considerably simplify the main arguments of the convergence analysis and avoid involved induction reasoning.

A third consequence of our approach is to provide a step-by-step analysis of our algorithm which singles out, at each stage of the convergence proof, the essential tools that are needed to get to the next stage. This allows one to understand the main ingredients

at play and to evidence the exact role of KL property in the analysis of algorithms in the nonconvex and nonsmooth setting, see more details in Section 3.2, where we outline a sort of "recipe" for proving global convergence results that could be of benefit to analyze many other optimization algorithms.

A fourth implication is that our block coordinate approach allows to get rid of a restrictive assumption inherent to the proximal forward-backward algorithm and which is often overlooked: the gradient of the smooth part H has to be *globally Lipschitz continuous*. This requirement often reduces the potential of applying PFB in concrete applications. On the contrary, our approach provides a flexibility that allows to deal with more general problems (*e.g.*, componentwise quadratic forms) or with some ill-conditioned quadratic problems. Indeed, the stepsizes in PALM may be adjusted componentwise in order to fit as much as possible the structure of the problem at hand, see Section 4 for an interesting application. Another by-product of this work is that it can also be applied to the convex version of Problem (M) for which convergence results are quite limited. Indeed, even for convex problems our convergence results are new (see the Appendix). Finally, to illustrate our results, we present a simple algorithm proven to converge to a critical point for a broad class of nonconvex and nonsmooth nonnegative matrix factorization problems, which to the best of our knowledge appears to be the first globally convergent algorithm for this important class of problems.

Outline of the paper. The paper is organized as follows. In the next section we define the problem, make precise our setting, and we collect a few preliminary basic facts on nonsmooth analysis, on proximal maps for nonconvex functions and we introduce the KL property. In Section 3 we state the algorithm PALM, derive some elementary properties and then develop a systematic approach to establish our main convergence results (see Section 3.2). In particular we clearly specify when and where the KL property is playing a fundamental role in the overall convergence analysis. Section 4 illustrates our results on a broad class of nonconvex and nonsmooth matrix factorization problems. Finally, to make this paper self-contained, we include an appendix which summarizes some well-known and relevant results on the KL property including some useful examples of KL functions. Throughout the paper, our notations are quite standard and can be found, for example, in [33].

2 The Problem and Some Preliminaries

2.1 The Problem and Basic Assumptions

We are interested in solving the nonconvex and nonsmooth minimization problem

$$(M) \qquad \text{minimize } \Psi(x,y) := f(x) + g(y) + H(x,y) \text{ over all } (x,y) \in \mathbb{R}^n \times \mathbb{R}^m.$$

Following [2], we take the following as our blanket assumption.

Assumption A. (i) $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^m \to (-\infty, +\infty]$ are proper and lower semicontinuous functions.

(ii) $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a C^1 function.

2.2 Subdifferentials of Nonconvex and Nonsmooth Functions

Let us recall few definitions concerning subdifferential calculus (see, for instance, [27, 33]). Recall that for $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ a proper and lower semicontinuous function, the domain of σ is defined through $\operatorname{dom} \sigma := \left\{x \in \mathbb{R}^d : \ \sigma(x) < +\infty\right\}$.

Definition 2.1 (Subdifferentials)**.** Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

(i) For a given $x \in \operatorname{dom} \sigma$, the *Fréchet subdifferential* of σ at x, written $\widehat{\partial}\sigma(x)$, is the set of all vectors $u \in \mathbb{R}^d$ which satisfy

$$\liminf_{y \neq x \ y \to x} \frac{\sigma(y) - \sigma(x) - \langle u, y - x\rangle}{\|y - x\|} \geq 0.$$

When $x \notin \operatorname{dom} \sigma$, we set $\widehat{\partial}\sigma(x) = \emptyset$.

(ii) The *limiting-subdifferential* [27], or simply the subdifferential, of σ at $x \in \mathbb{R}^n$, written $\partial\sigma(x)$, is defined through the following closure process

$$\partial\sigma(x) := \left\{u \in \mathbb{R}^d : \ \exists\, x^k \to x, \sigma\left(x^k\right) \to \sigma(x) \text{ and } u^k \in \widehat{\partial}\sigma\left(x^k\right) \to u \text{ as } k \to \infty\right\}.$$

Remark 2.1. (i) We have $\widehat{\partial}\sigma(x) \subset \partial\sigma(x)$ for each $x \in \mathbb{R}^d$. In the previous inclusion, the first set is closed and convex while the second one is closed (see [33, Theorem 8.6, page 302]).

(ii) Let $\left\{\left(x^k, u^k\right)\right\}_{k \in \mathbb{N}}$ be a sequence in $\operatorname{graph}(\partial\sigma)$ that converges to (x, u) as $k \to \infty$. By the very definition of $\partial\sigma(x)$, if $\sigma(x^k)$ converges to $\sigma(x)$ as $k \to \infty$, then $(x, u) \in \operatorname{graph}(\partial\sigma)$.

(iii) In this nonsmooth context, the well-known Fermat's rule remains barely unchanged. It formulates as: "if $x \in \mathbb{R}^d$ is a local minimizer of σ then $0 \in \partial\sigma(x)$".

(iv) Points whose subdifferential contains 0 are called *(limiting-)critical points*.

(v) The set of critical points of σ is denoted by $\operatorname{crit} \sigma$.

Definition 2.2 (Sublevel sets)**.** Being given real numbers α and β we set

$$[\alpha \leq \sigma \leq \beta] := \left\{x \in \mathbb{R}^d : \alpha \leq \sigma(x) \leq \beta\right\}.$$

We define similarly $[\alpha < \sigma < \beta]$. The level sets of σ are simply denoted by

$$[\sigma = \alpha] := \left\{x \in \mathbb{R}^d : \sigma(x) = \alpha\right\}.$$

Let us recall a useful result related to our structured Problem (M), see *e.g.*, [33].

Proposition 2.1 (Subdifferentiability property)**.** *Assume that the coupling function H in Problem (M) is continuously differentiable. Then for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ we have*

$$\partial\Psi(x,y) = (\nabla_x H(x,y) + \partial f(x), \nabla_y H(x,y) + \partial g(y)) = (\partial_x \Psi(x,y), \partial_y \Psi(x,y)). \quad (2.1)$$

Remark 2.2. Recall that for any set S, both $S + \emptyset$ and $S \times \emptyset$ are empty sets, so that the above formula makes sense over the whole product space $\mathbb{R}^n \times \mathbb{R}^m$.

2.3 Proximal Map for Nonconvex Functions

We need to recall the fundamental Moreau proximal map for a nonconvex function (see [33, page 20]). It is at the heart of the PALM algorithm.

Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lower semicontinuous function. Given $x \in \mathbb{R}^d$ and $t > 0$, the proximal map associated to σ and its corresponding Moreau proximal envelope are defined respectively by:

$$\operatorname{prox}_t^\sigma(x) := \operatorname{argmin}\left\{\sigma(u) + \frac{t}{2}\|u - x\|^2 : u \in \mathbb{R}^d\right\} \quad (2.2)$$

and

$$m^\sigma(x,t) := \inf\left\{\sigma(u) + \frac{1}{2t}\|u - x\|^2 : u \in \mathbb{R}^d\right\}.$$

Proposition 2.2 (Well-definedness of proximal maps)**.** *Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lower semicontinuous function with $\inf_{\mathbb{R}^d} \sigma > -\infty$. Then, for every $t \in (0, \infty)$ the set $\operatorname{prox}_{\frac{1}{t}}^\sigma(x)$ is nonempty and compact, in addition $m^\sigma(x,t)$ is finite and continuous in (x,t).*

Note that here $\operatorname{prox}_t^\sigma$ is a set-valued map. When $\sigma := \delta_X$, the indicator function of a nonempty and closed set X, *i.e.*, for the function $\delta_X : \mathbb{R}^d \to (-\infty, +\infty]$ defined, for all $x \in \mathbb{R}^d$, by

$$\delta_X(x) = \begin{cases} 0, & \text{if } x \in X, \\ +\infty, & \text{otherwise,} \end{cases}$$

the proximal map reduces to the projection operator onto X, defined by

$$P_X(v) := \operatorname{argmin}\{\|u - v\| : u \in X\}. \quad (2.3)$$

The projection $P_X : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ has nonempty values and defines in general a *multi-valued map*, as opposed to the convex case where orthogonal projections are guaranteed to be single-valued.

2.4 The Kurdyka-Łojasiewicz Property

The Kurdyka-Łojasiewicz property plays a central role in our analysis. Below, we recall the essential elements. We begin with the following extension of Łojasiewicz gradient inequality [25] as introduced in [2] for nonsmooth functions. First, we introduce some notation. For any subset $S \subset \mathbb{R}^d$ and any point $x \in \mathbb{R}^d$, the distance from x to S is defined and denoted by

$$\operatorname{dist}(x, S) := \inf \{\|y - x\| : \ y \in S\}.$$

When $S = \emptyset$, we have that $\operatorname{dist}(x, S) = \infty$ for all x.

Let $\eta \in (0, +\infty]$. We denote by Φ_η the class of all concave and continuous functions $\varphi : [0, \eta) \to \mathbb{R}_+$ which satisfy the following conditions

(i) $\varphi(0) = 0$;

(ii) φ is C^1 on $(0, \eta)$ and continuous at 0;

(iii) for all $s \in (0, \eta)$: $\varphi'(s) > 0$.

Now we define the Kurdyka-Łojasiewicz (KL) property.

Definition 2.3 (Kurdyka-Łojasiewicz property)**.** Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous.

(i) The function σ is said to have the *Kurdyka-Łojasiewicz (KL) property* at $\overline{u} \in \operatorname{dom} \partial\sigma := \left\{u \in \mathbb{R}^d : \partial\sigma(u) \neq \emptyset\right\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\overline{u}$ and a function $\varphi \in \Phi_\eta$, such that for all

$$u \in U \cap [\sigma(\overline{u}) < \sigma(u) < \sigma(\overline{u}) + \eta],$$

the following inequality holds

$$\varphi'(\sigma(u) - \sigma(\overline{u})) \operatorname{dist}(0, \partial\sigma(u)) \geq 1. \tag{2.4}$$

(ii) If σ satisfy the KL property at each point of $\operatorname{dom} \partial\sigma$ then σ is called a *KL function.*

It is easy to establish that KL property holds in the neighborhood of noncritical points (see, *e.g.*, [2]), thus the truly relevant aspect of this property is when $\bar{u}$ is critical, *i.e.*, when $0 \in \partial\sigma(\bar{u})$. In that case it warrants that σ is sharp up to a reparameterization of its values: "σ is amenable to sharpness". Indeed inequality (2.4) can be proved to imply

$$\operatorname{dist}(0, \partial(\varphi \circ (\sigma(u) - \sigma(\overline{u})))) \geq 1$$

for all convenient u (simply use the "one-sided" chain-rule [33, Theorem 10.6]). This means that the subgradients of the function $u \to \varphi \circ (\sigma(u) - \sigma(\bar{u}))$ have a norm greater than 1, no matter how close is the point u to the critical point $\bar{u}$ (provided that $\sigma(u) >$

$\sigma(\bar{u})$). This property is called *sharpness* while the reparameterization function φ is called a desingularizing function of σ at $\overline{u}$. As it is described further into detail, this geometrical feature has dramatic consequences in the study of first-order descent methods (see also [3]).

A remarkable aspect of KL functions is that they are ubiquitous in applications, for example, semi-algebraic, subanalytic and log-exp are KL functions (see [1, 2, 3] and references therein). These facts originates in the pioneering and fundamental works of Łojasiewicz [25] and Kurdyka [22]; works which were recently extended to nonsmooth functions in [16, 17]. In the Appendix we recall a nonsmooth semi-algebraic version of KL property, Theorem 5.1, which covers many problems arising in optimization and which plays a central role in the convergence analysis of our algorithm for the Nonnegative Matrix Factorization problem. For the reader's convenience, other related facts and pertinent results are also summarized in the same appendix.

3 PALM Algorithm and Convergence Analysis

3.1 The Algorithm PALM

As outlined in the Introduction, PALM can be viewed as alternating the steps of the PFB scheme. It is well-known that the proximal forward-backward scheme for minimizing the sum of a smooth function h with a nonsmooth one σ can simply be viewed as the proximal regularization of h linearized at a given point x^k, *i.e.*,

$$x^{k+1} \in \operatorname{argmin}_{x\in\mathbb{R}^d} \left\{ \left\langle x - x^k, \nabla h\left(x^k\right)\right\rangle + \frac{t}{2}\left\|x - x^k\right\|^2 + \sigma(x) \right\}, \quad (t>0), \tag{3.1}$$

that is, using the proximal map notation defined in (2.2), we get

$$x^{k+1} \in \operatorname{prox}_t^{\sigma}\left(x^k - \frac{1}{t}\nabla h\left(x^k\right)\right). \tag{3.2}$$

Adopting this scheme on Problem (M) we thus replace Ψ in the iterations (1.1) and (1.2) (*cf.* the Introduction) by their approximations which are obtained through the proximal linearization of each subproblems, *i.e.*, Ψ is replaced by

$$\widehat{\Psi}\left(x, y^k\right) = \left\langle x - x^k, \nabla_x H\left(x^k, y^k\right)\right\rangle + \frac{c_k}{2}\left\|x - x^k\right\|^2 + f(x), \quad (c_k > 0),$$

and

$$\widehat{\widehat{\Psi}}\left(x^{k+1}, y\right) = \left\langle y - y^k, \nabla_y H\left(x^{k+1}, y^k\right)\right\rangle + \frac{d_k}{2}\left\|y - y^k\right\|^2 + g(y), \quad (d_k > 0).$$

Thus alternating minimization on the two blocks (x, y) yields the basis of the algorithm PALM we propose here.

PALM: Proximal Alternating Linearized Minimization

1. Initialization: start with any $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$.
2. For each $k = 0, 1, \ldots$ generate a sequence $\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ as follows:

2.1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1\left(y^k\right)$ and compute

$$x^{k+1} \in \operatorname{prox}^f_{c_k}\left(x^k - \frac{1}{c_k}\nabla_x H\left(x^k, y^k\right)\right). \tag{3.3}$$

2.2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2\left(x^{k+1}\right)$ and compute

$$y^{k+1} \in \operatorname{prox}^g_{d_k}\left(y^k - \frac{1}{d_k}\nabla_y H\left(x^{k+1}, y^k\right)\right). \tag{3.4}$$

PALM needs minimal assumptions to be analyzed.

Assumption B. (i) $\inf_{\mathbb{R}^n\times\mathbb{R}^m} \Psi > -\infty$, $\inf_{\mathbb{R}^n} f > -\infty$ and $\inf_{\mathbb{R}^m} g > -\infty$.

(ii) For any fixed y the function $x \to H(x, y)$ is $C^{1,1}_{L_1(y)}$, namely the partial gradient $\nabla_x H(x, y)$ is globally Lipschitz with moduli $L_1(y)$, that is

$$\|\nabla_x H(x_1, y) - \nabla_x H(x_2, y)\| \le L_1(y)\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^n.$$

Likewise, for any fixed x the function $y \to H(x, y)$ is assumed to be $C^{1,1}_{L_2(x)}$.

(iii) For $i = 1, 2$ there exists $\lambda_i^-, \lambda_i^+ > 0$ such that

$$\inf\left\{L_1\left(y^k\right) : k \in \mathbb{N}\right\} \ge \lambda_1^- \quad \text{and} \quad \inf\left\{L_2\left(x^k\right) : k \in \mathbb{N}\right\} \ge \lambda_2^- \tag{3.5}$$

$$\sup\left\{L_1\left(y^k\right) : k \in \mathbb{N}\right\} \le \lambda_1^+ \quad \text{and} \quad \sup\left\{L_2\left(x^k\right) : k \in \mathbb{N}\right\} \le \lambda_2^+. \tag{3.6}$$

(iv) ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^n \times \mathbb{R}^m$. In other words, for each bounded subsets $B_1 \times B_2$ of $\mathbb{R}^n \times \mathbb{R}^m$ there exists $M > 0$ such that for all $(x_i, y_i) \in B_1 \times B_2$, $i = 1, 2$:

$$\begin{aligned}\|(\nabla_x H(x_1, y_1) - \nabla_x H(x_2, y_2), \nabla_y H(x_1, y_1) - \nabla_y H(x_2, y_2))\| \\ \le M\|(x_1 - x_2, y_1 - y_2)\|.\end{aligned} \tag{3.7}$$

A few words on Assumption B are now in order.

Remark 3.1. (i) Assumption B(i) ensures that Problem (M) is inf-bounded. It also warrants that the algorithm PALM is well defined through the proximal maps formulas (3.3) and (3.4) (see Proposition 2.2).

(ii) The partial Lipschitz properties required in Assumption B(ii) are at the heart of PALM which is designed to fully exploit the block-Lipschitz property of the problem at hand.

(iii) The inequalities (3.5) in Assumption B(iii) guarantees that the proximal steps in PALM remain always well-defined. As we describe below these two properties are not demanding at all.

Indeed, consider a function H whose gradient is Lipschitz continuous block-wise as in Assumption B(ii). Take now two arbitrary positive constants μ_1^- and μ_2^-, replace the Lipschitz modulis $L_1(y)$ and $L_2(x)$ by $L_1'(y) = \max\left\{L_1(y), \mu_1^-\right\}$ and $L_2'(x) = \max\left\{L_2(x), \mu_2^-\right\}$, respectively. The functions $L_1'(y)$ and $L_2'(x)$ are still some Lipschitz moduli of $\nabla_x H(\cdot, y)$ and $\nabla_y H(x, \cdot)$, respectively. Moreover

$$\inf\left\{L_1'(y) : y \in \mathbb{R}^m\right\} \geq \mu_1^- \quad \text{and} \quad \inf\left\{L_2'(x) : x \in \mathbb{R}^n\right\} \geq \mu_2^-.$$

Thus the inequalities (3.5) are trivially fulfilled with these new Lipschitz modulis and with $\lambda_i^- = \mu_i^-$ $(i = 1, 2)$.

(iv) Assumption B(iv) is satisfied whenever H is C^2 as a direct consequence of the Mean Value Theorem. Similarly, the inequalities (3.6) in Assumption B(iii), can be obtained by assuming that H is C^2 and that the generated sequence $\left\{\left(x^k, y^k\right)\right\}_{k\in\mathbb{N}}$ is bounded.

Before deriving the convergence results for PALM, in the next subsection we outline our proof methodology.

3.2 An Informal General Proof Recipe

Fix a positive integer N. Let $\Psi : \mathbb{R}^N \to (-\infty, +\infty]$ be a proper and lower semicontinuous function which is bounded from below and consider the problem

$$(P) \qquad \inf\left\{\Psi(z) : \ z \in \mathbb{R}^N\right\}.$$

Suppose we are given a generic algorithm $\mathcal{A}$ which generates a sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ via the following:

$$z^0 \in \mathbb{R}^N, z^{k+1} \in \mathcal{A}\left(z^k\right), \ k = 0, 1, \ldots.$$

The objective is to prove that the *whole sequence* generated by the algorithm $\mathcal{A}$ converges to a critical point of Ψ.

In the light of [1, 3], we outline a general methodology which describes the main steps to achieve this goal. In particular we put in evidence how and when the KL property is entering in action. Basically, the methodology consists of three main steps.

(i) **Sufficient decrease property:** Find a positive constant ρ_1 such that

$$\rho_1 \left\|z^{k+1} - z^k\right\|^2 \leq \Psi\left(z^k\right) - \Psi\left(z^{k+1}\right), \quad \forall k = 0, 1, \ldots.$$

(ii) **A subgradient lower bound for the iterates gap:** Assume that the sequence generated by the algorithm $\mathcal{A}$ is bounded.[1] Find another positive constant ρ_2, such that

$$\left\|w^{k+1}\right\| \leq \rho_2 \left\|z^{k+1} - z^k\right\|, \quad w^k \in \partial\Psi\left(z^k\right), \quad \forall k = 0, 1, \ldots.$$

These first two requirements above are quite standard and shared by essentially most *descent* algorithms, see *e.g.*, [2]. Note that when properties (i) and (ii) hold, then *for any* algorithm $\mathcal{A}$ one can show that the set of accumulations points is a nonempty, compact and connected set (see Lemma 3.5 (iii) for the case of PALM). One then need to prove that it is a subset of the critical points of Ψ on which Ψ is constant.

Apart from the aspects concerning the structure of the limiting set (nonempty, compact and connected), these first two steps depend on the structure of the specific chosen algorithm $\mathcal{A}$. Therefore the constants ρ_1 and ρ_2 are fit to the current given algorithm. The third step, needed to complete our goal, namely to establish *global convergence* to a critical point of Ψ, *doesn't depend at all* on the structure of the specific chosen algorithm $\mathcal{A}$.

Rather, it requires an additional assumption on the class of functions Ψ to be minimized. It is here that the KL property enters in action: relying on the descent property of the algorithm, and on a uniformization of the KL property (see Lemma 3.6) below, the third and last step amounts to:

(iii) **Using the KL property:** Assume that Ψ is a KL function and show that the generated sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a *Cauchy sequence*.

This basic approach can in principle be applied to any algorithm and is now systematically developed for PALM.

3.3 Basic Convergence Properties

We first establish some basic properties of PALM under our Assumptions A and B. We begin by recalling the well-known and important descent lemma for smooth functions, see *e.g.*, [12, 29].

Lemma 3.1 (Descent lemma)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with gradient ∇h assumed L_h-Lipschitz continuous. Then,*

$$h(u) \leq h(v) + \langle u - v, \nabla h(v)\rangle + \frac{L_h}{2}\|u - v\|^2, \ \forall\, u, v \in \mathbb{R}^d. \tag{3.8}$$

The main computational step of PALM involves a proximal map step of a proper and lower semicontinuous but nonconvex function. The next result shows that the well-known key inequality for the proximal-gradient step in the convex setting (see, *e.g.*, [9]) can be easily extended to the nonconvex setting to warrant sufficient decrease of the objective function after a proximal map step.

[1] For instance, it suffices to assume that Ψ is coercive to obtain a bounded sequence via (i); see also Remark 3.4.

Lemma 3.2 (Sufficient decrease property)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with gradient ∇h assumed L_h-Lipschitz continuous and let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function with $\inf_{\mathbb{R}^d} \sigma > -\infty$. Fix any $t > L_h$. Then, for any $u \in \operatorname{dom} \sigma$ and any $u^+ \in \mathbb{R}^d$ defined by*

$$u^+ \in \operatorname{prox}_t^\sigma \left(u - \frac{1}{t} \nabla h(u) \right) \tag{3.9}$$

we have

$$h\left(u^+\right) + \sigma\left(u^+\right) \leq h(u) + \sigma(u) - \frac{1}{2}(t - L_h)\left\|u^+ - u\right\|^2. \tag{3.10}$$

Proof. First, it follows immediately from Proposition 2.2 that u^+ is well-defined. By the definition of the proximal map given in (2.2) we get

$$u^+ \in \operatorname{argmin}_{v \in \mathbb{R}^d} \left\{ \langle v - u, \nabla h(u) \rangle + \frac{t}{2}\|v - u\|^2 + \sigma(v) \right\},$$

and hence in particular, taking $v = u$, we obtain

$$\left\langle u^+ - u, \nabla h(u) \right\rangle + \frac{t}{2}\left\|u^+ - u\right\|^2 + \sigma\left(u^+\right) \leq \sigma(u). \tag{3.11}$$

Invoking first the descent lemma (see Lemma 3.1) for h, and using then inequality (3.11), we get

$$\begin{aligned} h\left(u^+\right) + \sigma\left(u^+\right) &\leq h(u) + \left\langle u^+ - u, \nabla h(u) \right\rangle + \frac{L_h}{2}\left\|u^+ - u\right\|^2 + \sigma\left(u^+\right) \\ &\leq h(u) + \frac{L_h}{2}\left\|u^+ - u\right\|^2 + \sigma(u) - \frac{t}{2}\left\|u^+ - u\right\|^2 \\ &= h(u) + \sigma(u) - \frac{1}{2}(t - L_h)\left\|u^+ - u\right\|^2. \end{aligned}$$

This proves that (3.10) holds. □

Remark 3.2. (i) The above result is valid for any $t > 0$. The condition $t > L_h$ ensures a sufficient decrease in the value of $h(u^+) + \sigma(u^+)$.

(ii) If the function σ is taken as the indicator function δ_X of a nonempty, closed and nonconvex subset X, then the proximal map reduces to the projection P_X, that is

$$u^+ \in P_X \left(u - \frac{1}{t} \nabla h(u) \right)$$

and we recover the sufficient decrease property of the Projected Gradient Method (PGM) in the nonconvex case.

(iii) In the case when σ is a *convex*, proper and lower semicontinuous function, we can take $t = L_h$ (and even $t > \frac{L_h}{2}$). Indeed, in that case, we can apply the global optimality condition characterizing u^+ defined in (3.9) to get instead of (3.11) the stronger inequality

$$\sigma\left(u^+\right) + \left\langle u^+ - u, \nabla h\left(u\right)\right\rangle \leq \sigma\left(u\right) - t\left\|u^+ - u\right\|^2 \tag{3.12}$$

which together with the descent lemma (see Lemma 3.1) yields

$$h\left(u^+\right) + \sigma\left(u^+\right) \leq h\left(u\right) + \sigma\left(u\right) - \left(t - \frac{L_h}{2}\right)\left\|u^+ - u\right\|^2.$$

(iv) In view of item (iii), when applying PALM with convex functions f and g, the constants c_k and d_k, $k \in \mathbb{N}$, can simply be taken as $L_1\left(y^k\right)$ and $L_2\left(x^{k+1}\right)$, respectively.

Equipped with this result, we can now establish some useful properties for PALM under our Assumptions A and B. In the sequel for convenience we often use the notation

$$z^k := \left(x^k, y^k\right), \quad \forall k \geq 0.$$

Lemma 3.3 (Convergence properties)**.** *Suppose that Assumptions A and B hold. Let* $\left\{z^k\right\}_{k\in\mathbb{N}}$ *be a sequence generated by PALM. The following assertions hold.*

(i) *The sequence* $\left\{\Psi\left(z^k\right)\right\}_{k\in\mathbb{N}}$ *is nonincreasing and in particular*

$$\frac{\rho_1}{2}\left\|z^{k+1} - z^k\right\|^2 \leq \Psi\left(z^k\right) - \Psi\left(z^{k+1}\right), \quad \forall k \geq 0, \tag{3.13}$$

where

$$\rho_1 = \min\left\{\left(\gamma_1 - 1\right)\lambda_1^-, \left(\gamma_2 - 1\right)\lambda_2^-\right\}.$$

(ii) *We have*

$$\sum_{k=1}^{\infty}\left\|x^{k+1} - x^k\right\|^2 + \left\|y^{k+1} - y^k\right\|^2 = \sum_{k=1}^{\infty}\left\|z^{k+1} - z^k\right\|^2 < \infty, \tag{3.14}$$

and hence $\lim_{k\to\infty}\left\|z^{k+1} - z^k\right\| = 0$.

Proof. (i) Fix $k \geq 0$. Under our Assumption B(ii), the functions $x \to H\left(x, y\right)$ (y is fixed) and $y \to H\left(x, y\right)$ (x is fixed) are differentiable and have a Lipschitz gradient with modulis $L_1\left(y\right)$ and $L_2\left(x\right)$, respectively. Using the iterative steps (3.3) and (3.4), applying Lemma 3.2 twice, first with $h\left(\cdot\right) := H\left(\cdot, y^k\right)$, $\sigma := f$ and $t := c_k > L_1(y^k)$, and secondly with $h\left(\cdot\right) := H\left(x^{k+1}, \cdot\right)$, $\sigma := g$ and $t := d_k > L_2(x^{k+1})$, we obtain successively

$$\begin{aligned} H\left(x^{k+1}, y^k\right) + f\left(x^{k+1}\right) &\leq H\left(x^k, y^k\right) + f\left(x^k\right) - \frac{1}{2}\left(c_k - L_1\left(y^k\right)\right)\left\|x^{k+1} - x^k\right\|^2 \\ &= H\left(x^k, y^k\right) + f\left(x^k\right) - \frac{1}{2}\left(\gamma_1 - 1\right)L_1\left(y^k\right)\left\|x^{k+1} - x^k\right\|^2, \end{aligned}$$

and

$$H\left(x^{k+1},y^{k+1}\right)+g\left(y^{k+1}\right)\leq H\left(x^{k+1},y^{k}\right)+g\left(y^{k}\right)-\frac{1}{2}\left(d_k-L_2\left(x^{k+1}\right)\right)\left\|y^{k+1}-y^{k}\right\|^2$$
$$=H\left(x^{k+1},y^{k}\right)+g\left(y^{k}\right)-\frac{1}{2}\left(\gamma_2-1\right)L_2\left(x^{k+1}\right)\left\|y^{k+1}-y^{k}\right\|^2.$$

Adding the above two inequalities, we thus obtain for all $k\geq 0$,

$$\begin{aligned}&\Psi\left(z^{k}\right)-\Psi\left(z^{k+1}\right)\\&\quad=H\left(x^{k},y^{k}\right)+f\left(x^{k}\right)+g\left(y^{k}\right)-H\left(x^{k+1},y^{k+1}\right)-f\left(x^{k+1}\right)-g\left(y^{k+1}\right)\\&\quad\geq\frac{1}{2}\left(\gamma_1-1\right)L_1\left(y^{k}\right)\left\|x^{k+1}-x^{k}\right\|^2+\frac{1}{2}\left(\gamma_2-1\right)L_2\left(x^{k+1}\right)\left\|y^{k+1}-y^{k}\right\|^2.\end{aligned}\tag{3.15}$$

From (3.15) it follows that the sequence $\left\{\Psi\left(z^{k}\right)\right\}_{k\in\mathbb{N}}$ is nonincreasing, and since Ψ is assumed to be bounded from below (see Assumption B(i)), it converges to some real number $\underline{\Psi}$. Moreover, using the facts that $L_1\left(y^{k}\right)\geq\lambda_1^-\!>0$ and $L_2\left(x^{k+1}\right)\geq\lambda_2^-\!>0$ (see Assumption B(iii)), we get for all $k\geq 0$:

$$\begin{aligned}&\frac{1}{2}\left(\gamma_1-1\right)L_1\left(y^{k}\right)\left\|x^{k+1}-x^{k}\right\|^2+\frac{1}{2}\left(\gamma_2-1\right)L_2\left(x^{k+1}\right)\left\|y^{k+1}-y^{k}\right\|^2\\&\qquad\geq\frac{1}{2}\left(\gamma_1-1\right)\lambda_1^-\left\|x^{k+1}-x^{k}\right\|^2+\frac{1}{2}\left(\gamma_2-1\right)\lambda_2^-\left\|y^{k+1}-y^{k}\right\|^2\\&\qquad\geq\frac{\rho_1}{2}\left\|x^{k+1}-x^{k}\right\|^2+\frac{\rho_1}{2}\left\|y^{k+1}-y^{k}\right\|^2.\end{aligned}\tag{3.16}$$

Combining (3.15) and (3.16) yields the following

$$\frac{\rho_1}{2}\left\|z^{k+1}-z^{k}\right\|^2\leq\Psi\left(z^{k}\right)-\Psi\left(z^{k+1}\right),\tag{3.17}$$

and assertion (i) is proved.

(ii) Let N be a positive integer. Summing (3.17) from $k=0$ to $N-1$ we also get

$$\begin{aligned}\sum_{k=0}^{N-1}\left\|x^{k+1}-x^{k}\right\|^2+\left\|y^{k+1}-y^{k}\right\|^2&=\sum_{k=0}^{N-1}\left\|z^{k+1}-z^{k}\right\|^2\\&\leq\frac{2}{\rho_1}\left(\Psi\left(z^{0}\right)-\Psi\left(z^{N}\right)\right)\\&\leq\frac{2}{\rho_1}\left(\Psi\left(z^{0}\right)-\underline{\Psi}\right).\end{aligned}$$

Taking the limit as $N\to\infty$, we obtain the desired assertion (ii). □

3.4 Approaching the Set of Critical Points

In order to generate sequences approaching the set of critical points, we first prove the following result.

Lemma 3.4 (A subgradient lower bound for the iterates gap)**.** *Suppose that Assumptions A and B hold. Let* $\left\{z^k\right\}_{k\in\mathbb{N}}$ *be a sequence generated by PALM which is assumed to be bounded. For each positive integer* k*, define*

$$A_x^k := c_{k-1}\left(x^{k-1}-x^k\right)+\nabla_x H\left(x^k,y^k\right)-\nabla_x H\left(x^{k-1},y^{k-1}\right) \tag{3.18}$$

and

$$A_y^k := d_{k-1}\left(y^{k-1}-y^k\right)+\nabla_y H\left(x^k,y^k\right)-\nabla_y H\left(x^k,y^{k-1}\right). \tag{3.19}$$

Then $\left(A_x^k,A_y^k\right)\in\partial\Psi\left(x^k,y^k\right)$ *and there exists* $M>0$ *such that*

$$\left\|\left(A_x^k,A_y^k\right)\right\|\le\left\|A_x^k\right\|+\left\|A_y^k\right\|\le(2M+3\rho_2)\left\|z^k-z^{k-1}\right\|,\quad\forall k\ge 1, \tag{3.20}$$

where

$$\rho_2=\max\left\{\gamma_1\lambda_1^+,\gamma_2\lambda_2^+\right\}.$$

Proof. Let k be a positive integer. From the definition of the proximal map (2.2) and the iterative step (3.3) we have

$$x^k\in\operatorname{argmin}_{x\in\mathbb{R}^n}\left\{\left\langle x-x^{k-1},\nabla_x H\left(x^{k-1},y^{k-1}\right)\right\rangle+\frac{c_{k-1}}{2}\left\|x-x^{k-1}\right\|^2+f(x)\right\}.$$

Writing down the optimality condition yields

$$\nabla_x H\left(x^{k-1},y^{k-1}\right)+c_{k-1}\left(x^k-x^{k-1}\right)+u^k=0$$

where $u^k\in\partial f\left(x^k\right)$. Hence

$$\nabla_x H\left(x^{k-1},y^{k-1}\right)+u^k=c_{k-1}\left(x^{k-1}-x^k\right). \tag{3.21}$$

Similarly from the iterative step (3.4) we have

$$y^k\in\operatorname{argmin}_{y\in\mathbb{R}^m}\left\{\left\langle y-y^{k-1},\nabla_y H\left(x^k,y^{k-1}\right)\right\rangle+\frac{d_{k-1}}{2}\left\|y-y^{k-1}\right\|^2+g(y)\right\}.$$

Again, writing down the optimality condition yields

$$\nabla_y H\left(x^k,y^{k-1}\right)+d_{k-1}\left(y^k-y^{k-1}\right)+v^k=0$$

where $v^k\in\partial g\left(y^k\right)$. Hence

$$\nabla_y H\left(x^k,y^{k-1}\right)+v^k=d_{k-1}\left(y^{k-1}-y^k\right). \tag{3.22}$$

It is clear, from Proposition 2.1, that

$$\nabla_x H\left(x^k, y^k\right) + u^k \in \partial_x \Psi\left(x^k, y^k\right) \quad \text{and} \quad \nabla_y H\left(x^k, y^k\right) + v^k \in \partial_y \Psi\left(x^k, y^k\right).$$

From all these facts we obtain that $\left(A_x^k, A_y^k\right) \in \partial \Psi\left(x^k, y^k\right)$.

We now have to estimate the norms of A_x^k and A_y^k. Since ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^n \times \mathbb{R}^m$ (see Assumption B(iv)) and since we assumed that $\left\{z^k\right\}_{k\in\mathbb{N}}$ is bounded, there exists $M > 0$ such that

$$\begin{aligned}\left\|A_x^k\right\| &\leq c_{k-1}\left\|x^{k-1} - x^k\right\| + \left\|\nabla_x H\left(x^k, y^k\right) - \nabla_x H\left(x^{k-1}, y^{k-1}\right)\right\| \\ &\leq c_{k-1}\left\|x^k - x^{k-1}\right\| + M\left(\left\|x^k - x^{k-1}\right\| + \left\|y^k - y^{k-1}\right\|\right) \\ &= \left(M + c_{k-1}\right)\left\|x^k - x^{k-1}\right\| + M\left\|y^k - y^{k-1}\right\|.\end{aligned}$$

The moduli $L_1\left(y^{k-1}\right)$ being bounded from above by λ_1^+ (see Assumption B(iii)), we get that $c_{k-1} \leq \gamma_1 \lambda_1^+$ and thence

$$\begin{aligned}\left\|A_x^k\right\| &\leq \left(M + \gamma_1\lambda_1^+\right)\left\|x^k - x^{k-1}\right\| + M\left\|y^k - y^{k-1}\right\| \\ &\leq \left(2M + \gamma_1\lambda_1^+\right)\left\|z^k - z^{k-1}\right\| \\ &\leq \left(2M + \rho_2\right)\left\|z^k - z^{k-1}\right\|. \end{aligned} \tag{3.23}$$

On the other hand, from the Lipschitz continuity of $\nabla_y H\left(x, \cdot\right)$ (see Assumption B(ii)), we have that

$$\begin{aligned}\left\|A_y^k\right\| &\leq d_{k-1}\left\|y^k - y^{k-1}\right\| + \left\|\nabla_y H\left(x^k, y^k\right) - \nabla_y H\left(x^k, y^{k-1}\right)\right\| \\ &\leq d_{k-1}\left\|y^k - y^{k-1}\right\| + d_{k-1}\left\|y^k - y^{k-1}\right\| \\ &= 2d_{k-1}\left\|y^k - y^{k-1}\right\|.\end{aligned}$$

Since $L_2\left(x^k\right)$ is bounded from above by λ_2^+ (see Assumption B(iii)) we get that $d_{k-1} \leq \gamma_2\lambda_2^+$ and thence

$$\left\|A_y^k\right\| \leq 2\gamma_2\lambda_2^+\left\|y^k - y^{k-1}\right\| \leq 2\gamma_2\lambda_2^+\left\|z^k - z^{k-1}\right\| \leq 2\rho_2\left\|z^k - z^{k-1}\right\|. \tag{3.24}$$

Summing up these estimations, we get the desired result in (3.20), that is,

$$\left\|\left(A_x^k, A_y^k\right)\right\| \leq \left\|A_x^k\right\| + \left\|A_y^k\right\| \leq \left(2M + 3\rho_2\right)\left\|z^k - z^{k-1}\right\|.$$

This completes the proof. □

In the following result, we summarize several properties of the limit point set. Let $\left\{z^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by PALM from a starting point z^0. The set of all limit points is denoted by $\omega\left(z^0\right)$, *i.e.*,

$$\begin{aligned}\omega\left(z^0\right) = \big\{\overline{z} \in \mathbb{R}^n \times \mathbb{R}^m :\ &\exists \text{ an increasing sequence of integers } \left\{k_l\right\}_{l\in\mathbb{N}}, \\ &\text{such that } \ z^{k_l} \to \overline{z} \text{ as } l \to \infty\big\}.\end{aligned}$$

Lemma 3.5 (Properties of the limit point set $\omega(z^0)$)**.** *Suppose that Assumptions A and B hold. Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence generated by PALM which is assumed to be bounded. The following assertions hold.*

(i) $\emptyset \neq \omega(z^0) \subset \operatorname{crit}\Psi$

(ii) *We have*

$$\lim_{k\to\infty} \operatorname{dist}\left(z^k, \omega\left(z^0\right)\right) = 0. \tag{3.25}$$

(iii) *$\omega(z^0)$ is a nonempty, compact and connected set.*

(iv) *The objective function Ψ is finite and constant on $\omega(z^0)$.*

Proof. (i) Let $z^* = (x^*, y^*)$ be a limit point of $\{z^k\}_{k\in\mathbb{N}} = \{(x^k, y^k)\}_{k\in\mathbb{N}}$. This means that there is a subsequence $\{(x^{k_q}, y^{k_q})\}_{q\in\mathbb{N}}$ such that $(x^{k_q}, y^{k_q}) \to (x^*, y^*)$ as $q \to \infty$. Since f and g are lower semicontinuous (see Assumption A(i)), we obtain that

$$\liminf_{q\to\infty} f\left(x^{k_q}\right) \geq f(x^*) \quad \text{and} \quad \liminf_{q\to\infty} g\left(y^{k_q}\right) \geq g(y^*). \tag{3.26}$$

From the iterative step (3.3), we have for all integer k

$$x^{k+1} \in \operatorname{argmin}_{x\in\mathbb{R}^n} \left\{ \left\langle x - x^k, \nabla_x H\left(x^k, y^k\right)\right\rangle + \frac{c_k}{2}\left\|x - x^k\right\|^2 + f(x) \right\}.$$

Thus letting $x = x^*$ in the above, we get

$$\begin{aligned} &\left\langle x^{k+1} - x^k, \nabla_x H\left(x^k, y^k\right)\right\rangle + \frac{c_k}{2}\left\|x^{k+1} - x^k\right\|^2 + f\left(x^{k+1}\right) \\ &\qquad \leq \left\langle x^* - x^k, \nabla_x H\left(x^k, y^k\right)\right\rangle + \frac{c_k}{2}\left\|x^* - x^k\right\|^2 + f(x^*). \end{aligned}$$

Choosing $k = k_q - 1$ in the above inequality and letting q goes to ∞, we obtain

$$\begin{aligned} &\limsup_{q\to\infty} f\left(x^{k_q}\right) \\ &\qquad \leq \limsup_{q\to\infty} \left(\left\langle x^* - x^{k_q-1}, \nabla_x H\left(x^{k_q-1}, y^{k_q-1}\right)\right\rangle + \frac{c_k}{2}\left\|x^* - x^{k_q-1}\right\|^2 \right) + f(x^*), \end{aligned} \tag{3.27}$$

where we have used the facts that both sequences $\{x^k\}_{k\in\mathbb{N}}$ and $\{c_k\}_{k\in\mathbb{N}}$ are bounded, ∇H continuous and that the distance between two successive iterates tends to zero (see Lemma 3.3(ii)). For that very reason we also have $x^{k_q-1} \to x^*$ as $q \to \infty$, hence (3.27) reduces to $\limsup_{q\to\infty} f\left(x^{k_q}\right) \leq f(x^*)$. Thus, in view of (3.26), $f\left(x^{k_q}\right)$ tends to $f(x^*)$ as $q \to \infty$. Arguing similarly with g and y^k we thus finally obtain

$$\begin{aligned} \lim_{q\to\infty} \Psi\left(x^{k_q}, y^{k_q}\right) &= \lim_{q\to\infty} \left\{ H\left(x^{k_q}, y^{k_q}\right) + f\left(x^{k_q}\right) + g\left(y^{k_q}\right)\right\} \\ &= H(x^*, y^*) + f(x^*) + g(y^*) \\ &= \Psi(x^*, y^*). \end{aligned}$$

On the other hand we know from Lemmas 3.3(ii) and 3.4 that $\left(A_x^k, A_y^k\right) \in \partial\Psi\left(x^k, y^k\right)$ and $\left(A_x^k, A_y^k\right) \to (0,0)$ as $k \to \infty$. The closedness property of $\partial\Psi$ (see Remark 2.1(ii)) implies thus that $(0,0) \in \partial\Psi\left(x^*, y^*\right)$. This proves that (x^*, y^*) is a critical point of Ψ.

(ii) This item follows as an elementary consequence of the definition of limit points.

(iii) Set $\omega = \omega\left(z^0\right)$. Observe that ω can be viewed as an intersection of compact sets

$$\omega = \bigcap_{q \in \mathbb{N}} \overline{\bigcup_{k \geq q} \{z^k\}},$$

so it is also compact.

Towards a contradiction, we assume that ω is not connected. Whence there exist two nonempty and closed disjoint subsets A and B of ω such that $\omega = A \cup B$. Consider the function $\gamma : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$\gamma(z) = \frac{\operatorname{dist}(z, A)}{\operatorname{dist}(z, A) + \operatorname{dist}(z, B)}$$

for all $z \in \mathbb{R}^n \times \mathbb{R}^m$. Due to the closedness properties of A and B, the function γ is well defined, it is also continuous. Note that $A = \gamma^{-1}(\{0\}) = [\gamma = 0]$ and $B = \gamma^{-1}(\{1\}) = [\gamma = 1]$. Setting $U = [\gamma < 1/4]$ and $V = [\gamma > 3/4]$, we obtain, respectively, two open neighborhoods of the compact sets A and B. There exists an integer k_0 such that z^k either belongs to U or to V for all $k \geq k_0$. Supposing the contrary, there would exists a subsequence $\left\{z^{k_q}\right\}_{q \in \mathbb{N}}$ evolving in the complement of the open set $U \cup V$. This would imply the existence of a limit point z^* of z^k in $\mathbb{R}^n \setminus (U \cup V)$ which is impossible.

Put $r_k = \gamma\left(z^k\right)$ for each integer k. The sequence $\{r_k\}_{k \in \mathbb{N}}$ satisfies:

1. $r_k \notin [1/4, 3/4]$ for all $k \geq k_0$.

2. There exist infinitely many k such that $r_k < 1/4$.

3. There exist infinitely many k such that $r_k > 3/4$.

4. The difference $|r_{k+1} - r_k|$ tends to 0 as k goes to infinity.

The last point follows from the fact that γ is uniformly continuous on bounded sets together with the assumption that $\left\|z^{k+1} - z^k\right\| \to 0$. Clearly there exist no sequence complying with the above requirements. The set ω is therefore connected.

(iv) Denote by l the finite limit of $\Psi\left(z^k\right)$ as k goes to infinity. Take z^* in $\omega\left(z^0\right)$. There exists a subsequence z^{k_q} converging to z^* as q goes to infinity. On one hand the sequence $\left\{\Psi\left(z^{k_q}\right)\right\}_{q\in\mathbb{N}}$ converges to l and on the other hand (as we proved in assertion (i)) we have $\Psi\left(z^*\right)=l$. Hence the restriction of Ψ to $\omega\left(z^0\right)$ equals l. □

Remark 3.3. Note that properties (ii) and (iii) in Lemma 3.5 are generic for any sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ satisfying $\|z_{k+1}-z_k\|\to 0$ as k goes to infinity.

Our objective is now to prove that the sequence which is generated by PALM converges to a critical point of Problem (M). For that purpose we consider now that the objective of Problem (M) is a KL function, which is the case for example if f, g and H are semi-algebraic (see the Appendix for more details).

3.5 Convergence of PALM to Critical Points of Problem (M)

Before proving our main theorem the following result, which was established in [1, Lemma 1] for the Łojasiewicz property, would be adjusted within the more general KL property as follows.

Lemma 3.6 (Uniformized KL property)**.** *Let Ω be a compact set and let $\sigma:\mathbb{R}^d\to(-\infty,\infty]$ be a proper and lower semicontinuous function. Assume that σ is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon>0$, $\eta>0$ and $\varphi\in\Phi_\eta$ such that for all $\overline{u}$ in Ω and all u in the following intersection*

$$\left\{u\in\mathbb{R}^d:\ \operatorname{dist}\left(u,\Omega\right)<\varepsilon\right\}\cap\left[\sigma\left(\overline{u}\right)<\sigma\left(u\right)<\sigma\left(\overline{u}\right)+\eta\right] \tag{3.28}$$

one has,

$$\varphi'\left(\sigma\left(u\right)-\sigma\left(\overline{u}\right)\right)\operatorname{dist}\left(0,\partial\sigma\left(u\right)\right)\geq 1. \tag{3.29}$$

Proof. Denote by μ the value of σ over Ω. The compact set Ω can be covered by a finite number of open balls $B\left(u_i,\varepsilon_i\right)$ (with $u_i\in\Omega$ for $i=1,\ldots,p$) on which the KL property holds. For each $i=1,\ldots,p$, we denote the corresponding desingularizing function by $\varphi_i:[0,\eta_i)\to\mathbb{R}_+$ with $\eta_i>0$. For each $u\in B\left(u_i,\varepsilon_i\right)\cap[\mu<\sigma<\mu+\eta_i]$, we thus have

$$\varphi_i'\left(\sigma\left(u\right)-\sigma\left(u_i\right)\right)\operatorname{dist}\left(0,\partial\sigma\left(u\right)\right)=\varphi_i'\left(\sigma\left(u\right)-\mu\right)\operatorname{dist}\left(0,\partial\sigma\left(u\right)\right)\geq 1. \tag{3.30}$$

Choose $\varepsilon>0$ sufficiently small so that

$$U_\varepsilon:=\left\{x\in\mathbb{R}^d:\operatorname{dist}\left(x,\Omega\right)\leq\varepsilon\right\}\subset\bigcup_{i=1}^{p}B\left(u_i,\varepsilon_i\right). \tag{3.31}$$

Set $\eta=\min\left\{\eta_i:\ i=1,\ldots,p\right\}>0$ and

$$\varphi\left(s\right)=\sum_{i=1}^{p}\varphi_i\left(s\right),\ \forall s\in[0,\eta).$$

Observe now that, for all u in $U_\varepsilon \bigcap [\mu < \sigma < \mu + \eta]$, we obtain (*cf.* (3.30) and (3.31))

$$\varphi'(\sigma(u) - \mu)\,\mathrm{dist}(0, \partial\sigma(u)) = \sum_{i=1}^{p} \varphi_i'(\sigma(u) - \mu)\,\mathrm{dist}(0, \partial\sigma(u)) \geq 1.$$

This completes the proof. □

Now we will prove the main result.

Theorem 3.1 (A finite length property)**.** *Suppose that Ψ is a KL function such that Assumptions A and B hold. Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence generated by PALM which is assumed to be bounded. The following assertions hold.*

(i) *The sequence $\{z^k\}_{k\in\mathbb{N}}$ has finite length, that is,*

$$\sum_{k=1}^{\infty} \left\|z^{k+1} - z^k\right\| < \infty. \tag{3.32}$$

(ii) *The sequence $\{z^k\}_{k\in\mathbb{N}}$ converges to a critical point $z^* = (x^*, y^*)$ of Ψ.*

Proof. Since $\{z^k\}_{k\in\mathbb{N}}$ is bounded there exists a subsequence $\{z^{k_q}\}_{q\in\mathbb{N}}$ such that $z^{k_q} \to \overline{z}$ as $q \to \infty$. In a similar way as in Lemma 3.5(i) we get that

$$\lim_{k\to\infty} \Psi\left(x^k, y^k\right) = \Psi(\overline{x}, \overline{y}). \tag{3.33}$$

If there exists an integer $\bar{k}$ for which $\Psi\left(z^{\bar{k}}\right) = \Psi(\overline{z})$ then the decreasing property (3.13) would imply that $z^{\bar{k}+1} = z^{\bar{k}}$. A trivial induction show then that the sequence $\{z^k\}_{k\in\mathbb{N}}$ is stationary and the announced results are obvious. Since $\{\Psi(z^k)\}_{k\in\mathbb{N}}$ is a nonincreasing sequence, it is clear from (3.33) that $\Psi(\overline{z}) < \Psi(z^k)$ for all $k > 0$. Again from (3.33) for any $\eta > 0$ there exists a nonnegative integer k_0 such that $\Psi(z^k) < \Psi(\overline{z}) + \eta$ for all $k > k_0$. From (3.25) we know that $\lim_{k\to\infty} \mathrm{dist}\left(z^k, \omega(z^0)\right) = 0$. This means that for any $\varepsilon > 0$ there exists a positive integer k_1 such that $\mathrm{dist}\left(z^k, \omega(z^0)\right) < \varepsilon$ for all $k > k_1$. Summing up all these facts, we get that z^k belongs to the intersection in (3.28) for all $k > l := \max\{k_0, k_1\}$.

(i) Since $\omega(z^0)$ is nonempty and compact (see Lemma 3.5(ii)), and since Ψ is finite and constant on $\omega(z^0)$ (see Lemma 3.5(iv)), we can apply Lemma 3.6 with $\Omega = \omega(z^0)$. Therefore for any $k > l$ we have

$$\varphi'\left(\Psi\left(z^k\right) - \Psi(\overline{z})\right)\mathrm{dist}\left(0, \partial\Psi\left(z^k\right)\right) \geq 1. \tag{3.34}$$

This makes sense since we know that $\Psi\left(z^k\right) > \Psi(\overline{z})$ for any $k > l$. From Lemma 3.4 we get that

$$\varphi'\left(\Psi\left(z^k\right) - \Psi(\overline{z})\right) \geq \frac{1}{2M + 3\rho_2}\left\|z^k - z^{k-1}\right\|^{-1}. \tag{3.35}$$

On the other hand, from the concavity of φ we get that

$$\varphi\left(\Psi\left(z^k\right)-\Psi\left(\overline{z}\right)\right)-\varphi\left(\Psi\left(z^{k+1}\right)-\Psi\left(\overline{z}\right)\right) \\ \geq \varphi'\left(\Psi\left(z^k\right)-\Psi\left(\overline{z}\right)\right)\left(\Psi\left(z^k\right)-\Psi\left(z^{k+1}\right)\right). \quad (3.36)$$

For convenience, we define for all $p, q \in \mathbb{N}$ and $\overline{z}$ the following quantities

$$\Delta_{p,q} := \varphi\left(\Psi\left(z^p\right)-\Psi\left(\overline{z}\right)\right)-\varphi\left(\Psi\left(z^q\right)-\Psi\left(\overline{z}\right)\right),$$

and

$$C := \frac{2\left(2M+3\rho_2\right)}{\rho_1} \in (0,\infty).$$

Combining Lemma 3.3(i) with (3.35) and (3.36) yields for any $k > l$ that

$$\Delta_{k,k+1} \geq \frac{\left\|z^{k+1}-z^k\right\|^2}{C\left\|z^k-z^{k-1}\right\|}, \quad (3.37)$$

and hence

$$\left\|z^{k+1}-z^k\right\|^2 \leq C\Delta_{k,k+1}\left\|z^k-z^{k-1}\right\|.$$

Using the fact that $2\sqrt{\alpha\beta} \leq \alpha+\beta$ for all $\alpha, \beta \geq 0$, we infer

$$2\left\|z^{k+1}-z^k\right\| \leq \left\|z^k-z^{k-1}\right\| + C\Delta_{k,k+1}. \quad (3.38)$$

Let us now prove that for any $k > l$ the following inequality holds

$$\sum_{i=l+1}^{k}\left\|z^{i+1}-z^i\right\| \leq \left\|z^{l+1}-z^l\right\| + C\Delta_{l+1,k+1}.$$

Summing up (3.38) for $i = l+1, \ldots, k$ yields

$$\begin{aligned} 2\sum_{i=l+1}^{k}\left\|z^{i+1}-z^i\right\| &\leq \sum_{i=l+1}^{k}\left\|z^i-z^{i-1}\right\| + C\sum_{i=l+1}^{k}\Delta_{i,i+1} \\ &\leq \sum_{i=l+1}^{k}\left\|z^{i+1}-z^i\right\| + \left\|z^{l+1}-z^l\right\| + C\sum_{i=l+1}^{k}\Delta_{i,i+1} \\ &= \sum_{i=l+1}^{k}\left\|z^{i+1}-z^i\right\| + \left\|z^{l+1}-z^l\right\| + C\Delta_{l+1,k+1} \end{aligned}$$

where the last inequality follows from the fact that $\Delta_{p,q}+\Delta_{q,r} = \Delta_{p,r}$ for all $p, q, r \in \mathbb{N}$. Since $\varphi \geq 0$, we thus have for any $k > l$ that

$$\sum_{i=l+1}^{k}\left\|z^{i+1}-z^i\right\| \leq \left\|z^{l+1}-z^l\right\| + C\varphi\left(\Psi\left(z^{l+1}\right)-\Psi\left(\overline{z}\right)\right).$$

This easily shows that the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ has finite length, that is,

$$\sum_{k=1}^{\infty}\left\|z^{k+1}-z^k\right\|<\infty. \tag{3.39}$$

(ii) It is clear that (3.39) implies that the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a Cauchy sequence and hence is a convergent sequence. Indeed, with $q>p>l$ we have

$$z^q-z^p=\sum_{k=p}^{q-1}\left(z^{k+1}-z^k\right)$$

hence

$$\left\|z^q-z^p\right\|=\left\|\sum_{k=p}^{q-1}\left(z^{k+1}-z^k\right)\right\|\leq\sum_{k=p}^{q-1}\left\|z^{k+1}-z^k\right\|.$$

Since (3.39) implies that $\sum_{k=l+1}^{\infty}\left\|z^{k+1}-z^k\right\|$ converges to zero as $l\to\infty$, it follows that $\left\{z^k\right\}_{k\in\mathbb{N}}$ is a Cauchy sequence and hence is a convergent sequence. Now the result follows immediately from Lemma 3.5(i).

This completes the proof. □

Remark 3.4. (i) The boundedness assumption on the generated sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ holds in several scenarios such as when the functions f and g have bounded level sets. For a few more scenarios see [2].

(ii) An important and fundamental case of application of Theorem 3.1 is when the data functions f, g and H are semi-algebraic. Observe also that the desingularizing function for semi-algebraic problems can be chosen to be of the form

$$\varphi\left(s\right)=cs^{1-\theta}, \tag{3.40}$$

where c is positive real number and θ belongs to $[0,1)$ (see [1] for more details). As explained below, this fact impacts the convergence rate of the method.

If the desingularizing function φ of Ψ is of the form (3.40), then, as in [1] the following estimations hold.

(i) If $\theta=0$ then the sequence $\left\{z^k\right\}_{k\in\mathbb{N}}$ converges in a finite number of steps.

(ii) If $\theta\in(0,1/2]$ then there exist $\omega>0$ and $\tau\in[0,1)$ such that $\left\|z^k-\overline{z}\right\|\leq\omega\,\tau^k$.

(iii) If $\theta\in(1/2,1)$ then there exist $\omega>0$ such that

$$\left\|z^k-\overline{z}\right\|\leq\omega\,k^{-\frac{1-\theta}{2\theta-1}}.$$

3.6 Extension of PALM for p Blocks

The simple structure of PALM allows to extend it to the more general setting involving $p > 2$ blocks for which Theorem 3.1 holds. This is briefly outlined below. Suppose that our optimization problem is now given as

$$\text{minimize}\left\{\Psi\left(x_1,\ldots,x_p\right) := \sum_{i=1}^{p} f_i\left(x_i\right) + H\left(x_1,\ldots,x_p\right) :\ x_i \in \mathbb{R}^{n_i}\right\},$$

where $H : \mathbb{R}^N \to \mathbb{R}$ with $N = \sum_{i=1}^{p} n_i$ is assumed to be C^1 and each f_i, $i = 1,\ldots,p$, is a proper and lower-semicontinuous function (this is exactly Assumption A for $p > 2$). We also assume that a modified version of Assumption B for $p > 2$ blocks holds. In this case we denote by $\nabla_i H$ the gradient of H with respect to variable x_i, $i = 1,\ldots,p$. We denote by L_i, $i = 1,\ldots,p$, the Lipschitz moduli of $\nabla_i H\left(x_1,\ldots,\cdot,\ldots,x_p\right)$, that is, the gradient of H with respect to variable x_i when all x_j, $i \neq j$ $(j = 1,\ldots,p)$, are fixed. Similarly to Assumption B(ii), it is clear that each L_i, $i = 1,\ldots,p$, is a function of the $p-1$ variables x_j, $j \neq i$ $(j = 1,\ldots,p)$.

For simplicity of the presentation of PALM for the case of $p > 2$ blocks we will use the following notations. Denote $x^k = \left(x_1^k, x_2^k, \ldots, x_p^k\right)$ and

$$x^k(i) = \left(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_p^k\right).$$

Therefore $x^k(0) = \left(x_1^k, x_2^k \ldots, x_p^k\right) = x^k$ and $x^k(p) = \left(x_1^{k+1}, x_2^{k+1}, \ldots, x_p^{k+1}\right) = x^{k+1}$.

In this case the algorithm PALM minimizes Ψ with respect to each $x_1,\ldots,x_p$, taken in cyclic order while fixing the previous computed iterate. More precisely, starting with any $\left(x_1^0, x_2^0, \ldots, x_p^0\right) \in \mathbb{R}^N$, PALM generates a sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ via the following successively scheme:

$$x_i^{k+1} \in \operatorname{prox}_{c_i^k}^{f_i}\left(x_i^k - \frac{1}{c_i^k}\nabla_i H\left(x^k(i-1)\right)\right), \quad i = 1,2,\ldots,p,$$

where $c_i^k = \gamma_i L_i$ and $\gamma_i > 1$. Theorem 3.1 can then be applied for the p-blocks version of PALM.

3.7 The Proximal Forward-Backward Scheme

When there is no y term, PALM reduces to PFB. In this case we have $\Psi\left(x\right) := f\left(x\right) + h\left(x\right)$ (where $h\left(x\right) \equiv H\left(x,0\right)$), and the proximal forward-backward scheme for minimizing Ψ can simply be viewed as the proximal regularization of h linearized at a given point x^k, *i.e.*,

$$x^{k+1} \in \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{\left\langle x - x^k, \nabla h\left(x^k\right)\right\rangle + \frac{t_k}{2}\left\|x - x^k\right\|^2 + f\left(x\right)\right\}.$$

A convergence result for the PFB scheme was first proved in [3] via the abstract framework developed in that paper. Our approach allows for a simpler and more direct proof. The

sufficient decrease property of the sequence $\left\{\Psi\left(x^k\right)\right\}_{k\in\mathbb{N}}$ follows directly from Lemma 3.2 with $\sigma := f$ and $t := t_k > L_h$. The second property "a subgradient lower bound for the iterates gap" follows from the Lipschitz continuity of ∇h. Now the globally convergent result follows immediately from Theorem 3.1. For the sake of completeness we record the result in the following proposition.

Proposition 3.1 (A convergence result of PFB)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with gradient ∇h assumed L_h-Lipschitz continuous and let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function with $\inf_{\mathbb{R}^d} f > -\infty$. Assume that $f + h$ is a KL function. Let $\left\{x^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by PFB which is assumed to be bounded and let $t_k > L_h$. The following assertions hold.*

(i) *The sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ has finite length, that is,*

$$\sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\| < \infty.$$

(ii) *The sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ converges to a critical point x^* of $f + h$.*

It is well-known that PFB reduces to the projected gradient method (PGM) when $f = \delta_X$ (where X is a nonempty, closed and nonconvex subset of $\mathbb{R}^d$), *i.e.*, PGM generates a sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ via

$$x^{k+1} \in P_X\left(x^k - \frac{1}{t_k}\nabla h\left(x^k\right)\right).$$

Thus when $h+\delta_X$ is a KL function and $h \in C^{1,1}_{L_h}$, global convergence of the sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ generated by PGM follows from Proposition 3.1, and recovers the result established in [3].

4 An Application to Matrix Factorization Problems

Matrix factorization problems play a fundamental role in data analysis and can be found in many disparate applications. A very large body of literature covers this active research area; for a recent account we refer for example to the book [18] and references therein.

In this section we show how PALM can be applied to a broad class of such problems to produce a globally convergent algorithm.

4.1 A Broad Class of Matrix Factorization Problems

Let p, q, m, n and r be given integers. Define the following sets in the space of real matrices

$$\begin{aligned}
\mathcal{K}_{p,q} &= \left\{M \in \mathbb{R}^{p\times q} :\ M \geq 0\right\}, \\
\mathcal{F} &= \left\{X \in \mathbb{R}^{m\times r} :\ R_1(X) \leq \alpha\right\}, \\
\mathcal{G} &= \left\{Y \in \mathbb{R}^{r\times n} :\ R_2(Y) \leq \beta\right\},
\end{aligned}$$

where R_1 and R_2 are lower semicontinuous functions and $\alpha, \beta \in \mathbb{R}_+$ are given parameters.

Roughly speaking, the matrix factorization (or approximation) problem consists in finding a product decomposition of a given matrix satisfying certain properties.

The Problem

Given a matrix $A \in \mathbb{R}^{m\times n}$ and let r be an integer which is much smaller than $\min\{m, n\}$, find two matrices $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{r\times n}$ such that

$$\begin{cases} A \approx XY, \\ X \in \mathcal{K}_{m,r} \cap \mathcal{F}, \\ Y \in \mathcal{K}_{r,n} \cap \mathcal{G}. \end{cases}$$

The functions R_1 and R_2 are often used to describe some additional features of the matrices X and Y, respectively, arising in a specific application at hand (see more below).

To solve the problem, we adopt the optimization approach, that is, we consider the nonconvex and nonsmooth minimization problem

$$(MF) \qquad \min\left\{d\left(A, XY\right) :\ X \in \mathcal{K}_{m,r} \cap \mathcal{F}, Y \in \mathcal{K}_{r,n} \cap \mathcal{G}\right\},$$

where $d : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}_+$ stands as a proximity function measuring the quality of the approximation, satisfying $d\left(U, V\right) = 0$ if and only if $U = V$. Note that $d\left(\cdot, \cdot\right)$ is not necessarily symmetric and is not a metric.

Another way to formulate (MF) is to consider its penalized version where the "hard" constraints are the candidates to be penalized, *i.e.*, we consider the following penalized problem

$$(P - MF) \qquad \min\left\{\mu_1 R_1\left(X\right) + \mu_2 R_2\left(Y\right) + d\left(A, XY\right) : X \in \mathcal{K}_{m,r}, Y \in \mathcal{K}_{r,n}\right\},$$

where μ_1 and $\mu_2 > 0$ are penalty parameters. However, note that the penalty approach requires the tuning of the unknown penalty parameters which might be a difficult issue.

Both formulations can be written in the form of our general Problem (M) with the obvious identifications for the corresponding H, f and g, *e.g.*,

$$\min\left\{\Psi\left(X, Y\right) := f\left(X\right) + g\left(Y\right) + H\left(X, Y\right) :\ X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{r\times n}\right\},$$

where

$$\text{MF-Constrained } \Psi_c\left(X, Y\right) := \delta_{\mathcal{K}_{m,r}\cap\mathcal{F}}\left(X\right) + \delta_{\mathcal{K}_{r,n}\cap\mathcal{G}}\left(Y\right) + d\left(A, XY\right),$$
$$\text{MF-Penalized } \Psi_p\left(X, Y\right) := \mu_1 R_1\left(X\right) + \delta_{\mathcal{K}_{m,r}}\left(X\right) + \mu_2 R_2\left(Y\right) + \delta_{\mathcal{K}_{r,n}}\left(Y\right) + d\left(A, XY\right).$$

Thus, assuming that Assumptions A and B hold for the problem data quantified here via $[d, \mathcal{F}, \mathcal{G}]$, and that the functions d, R_1 and R_2 are KL functions, we can apply PALM and Theorem 3.1 to produce a globally convergent scheme to a critical point of Ψ that solves the (MF) problem. The above does not seem to have been addressed in the literature within such a general formalism. It covers a multitude of possible formulations from which many algorithms can be conceived by appropriate choices of the triple $[d, \mathcal{F}, \mathcal{G}]$ within a given application at hands. This is illustrated next on an important class of problems.

4.2 An Algorithm for the Sparse Nonnegative Matrix Factorization

To be specific, in the sequel we focus on the classical case where the proximity measure is defined via the Frobenius norm

$$d(A, XY) = \frac{1}{2} \|A - XY\|_F^2$$

and for any matrix M, the Frobenius norm is defined by

$$\|M\|_F^2 = \sum_{i,j} m_{ij}^2 = \mathrm{Tr}\left(MM^T\right) = \mathrm{Tr}\left(M^T M\right) = \langle M, M \rangle,$$

where Tr is the Trace operator. Many other proximity measures can also be used, such as entropy-like distances, see *e.g.*, [18] and references therein.

Example 4.1 (Nonnegative matrix factorization)**.** With $\mathcal{F} = \mathbb{R}^{m \times r}$ and $\mathcal{G} = \mathbb{R}^{r \times n}$, the Problem (MF) reduces to the so called Nonnegative Matrix Factorization (NMF) problem

$$\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : X \geq 0, Y \geq 0 \right\}.$$

The nonnegative matrix factorization [23] has been at the heart of intense research applied to a variety of applications (see, *e.g.*, [14] for applications in signal processing). More recently the introduction of "sparsity" has been of particular importance, and variants of NMF involving sparsity has also been considered in the literature (see, *e.g.*, [20, 21]). Many, if not most, algorithms are based on the Gauss-Seidel like method for solving the NMF problem, see *e.g.*, [11, 18, 24], and with quite limited convergence results. Moreover, extended versions of NMF with sparsity were considered via relaxations and corresponding convex re-formulations solved by sophisticated and computationally demanding conic programming schemes, see *e.g.*, [20, 21].

To illustrate the benefit of our approach, we now show how PALM can be applied to solve directly the more difficult constrained nonconvex and nonsmooth sparse nonnegative matrix factorization problem "as is", and produces a simple convergent scheme.

First we note that the objective function $d(A, XY) := H(X, Y) = (1/2) \|A - XY\|_F^2$ is a real polynomial function hence semi-algebraic; moreover, both functions $X \to H(X, Y)$ (for fixed Y) and $Y \to H(X, Y)$ (for fixed X), are $C^{1,1}$. Indeed we have

$$X \to \nabla_X H(X, Y) = (XY - A) Y^T \quad \text{and} \quad Y \to \nabla_Y H(X, Y) = X^T (XY - A)$$

which are Lipschitz continuous with $L_1(Y) \equiv \left\|YY^T\right\|_F$ and $L_2(X) \equiv \left\|X^T X\right\|_F$ as Lipschitz modulis, respectively.

As a specific case, let us now consider the overall sparsity measure of a matrix defined by

$$R_1(X) = \|X\|_0 := \sum_i \|x_i\|_0, \ (x_i \text{ column vector of } X)$$

which counts the number of nonzero elements in the matrix X. Similarly $R_2(Y) = \|Y\|_0$.

As shown in Example 5.2 (see the Appendix) both functions R_1 and R_2 are semi-algebraic. Thanks to the properties of semi-algebraic functions (see the Appendix) it follows that Ψ_c is semi-algebraic and PALM could be applied to produce a globally convergent algorithm. However, to apply PALM properly, we need to compute the proximal map of the nonconvex function $\|X\|_0$ on $X \geq 0$ for some given matrix U. It turns out that this can be done effectively, as the next proposition shows. Our result makes use of the following operator (see, *e.g.*, [26]).

Definition 4.1. Given any matrix $U \in \mathbb{R}^{m\times n}$, define the operator $T_s : \mathbb{R}^{m\times n} \rightrightarrows \mathbb{R}^{m\times n}$ by

$$T_s(U) := \operatorname{argmin}_{V\in\mathbb{R}^{m\times n}} \left\{ \|U - V\|_F^2 : \ \|V\|_0 \leq s \right\}.$$

Observe that the operator T_s is in general multi-valued. For a given matrix U, it is actually easy to see that the elements of $T_s(U)$ are obtained by choosing exactly s indices corresponding the s first largest entries (in absolute value) of U and by setting $(T_s(U))_{ij} = U_{ij}$ for such indices and $(T_s(U))_{ij} = 0$ otherwise. The multi-valuedness of T_s comes from the fact that the s largest entries may not be uniquely defined.

Since computing T_s only requires determining the s^{th} largest numbers of a matrix of mn numbers, this can be done in $\mathcal{O}(mn)$ time [13] and zeroing out the proper entries in one more pass of the mn numbers.

We define the usual projection map onto $\mathbb{R}^{m\times n}_+$ by

$$P_+(U) := \operatorname{argmin}_{V\in\mathbb{R}^{m\times n}} \left\{ \|U - V\|_F^2 : \ V \geq 0 \right\} = \max\{0, U\},$$

where the max operation is taken componentwise.

Proposition 4.1 (Proximal map formula)**.** *Let $U \in \mathbb{R}^{m\times n}$ and let $f := \delta_{X\geq 0} + \delta_{\|X\|_0\leq s}$. Then*

$$\operatorname{prox}_1^f(U) = \operatorname{argmin}\left\{ \frac{1}{2}\|X - U\|_F^2 : X \geq 0, \|X\|_0 \leq s \right\} = T_s(P_+(U))$$

where T_s is defined in Definition 4.1.

Proof. Given any matrix $U \in \mathbb{R}^{m\times n}$, let us introduce the following notations

$$\|X\|_+^2 = \sum_{(i,j)\in\mathcal{I}^+} X_{ij}^2 \quad \text{and} \quad \|X\|_-^2 = \sum_{(i,j)\in\mathcal{I}^-} X_{ij}^2,$$

where

$$\mathcal{I}^+ = \{(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\} : \ U_{ij} \geq 0\}$$

and

$$\mathcal{I}^- = \{(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\} : \ U_{ij} < 0\}.$$

Observe that the following relations hold

$$(i) \quad \|X\|_F^2 = \|X\|_+^2 + \|X\|_-^2 \qquad (ii) \quad \|X - U\|_+^2 + \|X\|_-^2 = \|X - P_+(U)\|_F^2$$

and

$$(iii) \quad \|X\|_-^2 = 0 \Leftrightarrow X_{ij} = 0 \quad \forall\, (i,j) \in \mathcal{I}^-,$$

where the second relation follows from relation (i) and the fact that $(P_+(U))_{ij} = U_{ij}$ for any $(i,j) \in \mathcal{I}^+$ and $(P_+(U))_{ij} = 0$ for any $(i,j) \in \mathcal{I}^-$.

From the above fact (i), we thus have that $\bar{X} \in \mathrm{prox}_1^f(U)$ if and only if

$$\begin{aligned}
\bar{X} &\in \operatorname{argmin} \left\{ \|X-U\|_F^2 :\ X \geq 0,\ \|X\|_0 \leq s \right\} \\
&= \operatorname{argmin} \left\{ \|X-U\|_+^2 + \|X-U\|_-^2 :\ X \geq 0,\ \|X\|_0 \leq s \right\} \\
&= \operatorname{argmin} \left\{ \|X-U\|_+^2 + \|X\|_-^2 - 2 \sum_{(i,j)\in\mathcal{I}^-} X_{ij}U_{ij} :\ X \geq 0,\ \|X\|_0 \leq s \right\} \qquad (4.1) \\
&= \operatorname{argmin} \left\{ \|X-U\|_+^2 :\ X_{ij} = 0 \quad \forall\, (i,j) \in \mathcal{I}^-,\ X \geq 0,\ \|X\|_0 \leq s \right\}, \qquad (4.2)
\end{aligned}$$

where the last equality follows from the fact that every solution of (4.2) is clearly a solution of (4.1), while the converse implication follows by a simple contradiction argument. Arguing in a similar way, one can see that the constraint $X \geq 0$ in problem (4.2) can be removed without affecting the optimal solution of that problem. Thus, recalling the facts (ii) and (iii) we obtain

$$\begin{aligned}
\bar{X} &\in \operatorname{argmin} \left\{ \|X-U\|_+^2 : \|X\|_-^2 = 0,\ \|X\|_0 \leq s \right\} \\
&= \operatorname{argmin} \left\{ \|X-U\|_+^2 + \|X\|_-^2 :\ \|X\|_0 \leq s \right\} \\
&= \operatorname{argmin} \left\{ \|X-P_+(U)\|_F^2 :\ \|X\|_0 \leq s \right\} = T_s\left(P_+(U)\right),
\end{aligned}$$

where the last equality is by the definition of T_s (see Definition 4.1). □

With $R_1 := \delta_{X\geq 0} + \delta_{\|X\|_0 \leq \alpha}$ and $R_2 := \delta_{Y\geq 0} + \delta_{\|Y\|_0 \leq \beta}$, we now have all the ingredients to apply PALM and formulate explicitly a simple algorithm for the sparse nonnegative matrix factorization problem.

PALM-Sparse NMF

1. Initialization: Select random nonnegative matrices $X^0 \in \mathbb{R}^{m\times r}$ and $Y^0 \in \mathbb{R}^{r\times n}$.
2. For each $k = 0, 1, \ldots$ generate a sequence $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ as follows:

2.1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 \left\|Y^k \left(Y^k\right)^T\right\|_F$ and compute

$$\begin{aligned} U^k &= X^k - \frac{1}{c_k}\left(X^k Y^k - A\right)\left(Y^k\right)^T, \\ X^{k+1} &\in \operatorname{prox}_{c_k}^{R_1}\left(U^k\right) = T_\alpha\left(P_+\left(U^k\right)\right). \end{aligned} \tag{4.3}$$

2.2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 \left\|X^{k+1}\left(X^{k+1}\right)^T\right\|_F$ and compute

$$\begin{aligned} V^k &= Y^k - \frac{1}{d_k}\left(X^{k+1}\right)^T\left(X^{k+1}Y^k - A\right), \\ Y^{k+1} &\in \operatorname{prox}_{d_k}^{R_2}\left(V^k\right) = T_\beta\left(P_+\left(V^k\right)\right). \end{aligned}$$

Remark 4.1. (i) Observe that PALM-Sparse NMF requires that the Lipschitz modulis $\left\|X^{k+1}\left(X^{k+1}\right)^T\right\|_F$ and $\left\|Y^k\left(Y^k\right)^T\right\|_F$ remain bounded away from zero. This means equivalently that we assume that

$$\inf_{k\in\mathbb{N}}\left\{\left\|X^k\right\|_F, \left\|Y^k\right\|_F\right\} > 0.$$

In view of Remark 3.1(iii), we could avoid this assumption by introducing a safeguard $\nu > 0$ and simply replacing the Lipschitz modulis in PALM-Sparse NMF by

$$\max\left(\nu, \left\|X^{k+1}\left(X^{k+1}\right)^T\right\|_F\right) \quad \text{and} \quad \max\left(\nu, \left\|Y^k\left(Y^k\right)^T\right\|_F\right).$$

(ii) Note that the easier nonnegative matrix factorization problem given in Example 4.1 is a particular instance of the sparse NMF and in that case both operators T_α and T_β reduce to the identity operators. Hence, the computation in Step 2.1. for NMF reduces to

$$X^{k+1} = P_+\left(U^k\right)$$

where U^k is given in (4.3) (similarly for Y^{k+1}). Moreover, since in that case the constraints set $\mathcal{K}_{m,r}$ and $\mathcal{K}_{r,n}$ are closed and convex, it follows from Remark 3.2(iii) that we can set $c_k = \left\|Y^k\left(Y^k\right)^T\right\|_F$ and $d_k = \left\|X^{k+1}\left(X^{k+1}\right)^T\right\|_F$ in that case.

The assumptions required to apply PALM are clearly satisfied and hence we can use Theorem 3.1 in order to obtain that the generated sequence is globally convergent to a critical point of the Sparse NMF problem (and similarly for NMF, as a special case). We record this in the following theorem.

Theorem 4.1. *Let* $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ *be a sequence generated by PALM-Sparse NMF which is assumed to be bounded and to satisfy* $\inf_{k\in\mathbb{N}}\left\{\left\|X^k\right\|_F, \left\|Y^k\right\|_F\right\} > 0$. *Then,*

(i) *The sequence* $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ *has finite length, that is*

$$\sum_{k=1}^{\infty} \left\|X^{k+1} - X^k\right\|_F + \left\|Y^{k+1} - Y^k\right\|_F < \infty.$$

(ii) *The sequence* $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ *converges to a critical point* (X^*, Y^*) *of the Sparse NMF.*

5 Appendix: KL Results

This appendix summarizes some important results on KL theory and gives some examples.

Definition 5.1 (Semi-algebraic sets and functions)**.** (i) A subset S of $\mathbb{R}^d$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^d \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{u \in \mathbb{R}^d : \ g_{ij}(u) = 0 \text{ and } \ h_{ij}(u) < 0\right\}.$$

(ii) A function $h : \mathbb{R}^d \to (-\infty, +\infty]$ is called semi-algebraic if its graph

$$\left\{(u, t) \in \mathbb{R}^{d+1} : \ h(u) = t\right\}$$

is a semi-algebraic subset of $\mathbb{R}^{d+1}$.

The following result is a nonsmooth version of the Łojasiewicz gradient inequality, it can be found in [16, 17].

Theorem 5.1. *Let* $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ *be a proper and lower semicontinuous function. If* σ *is semi-algebraic then it satisfies the KL property at any point of* $\operatorname{dom}\sigma$.

The class of semi-algebraic sets is stable under the following operations: finite unions, finite intersections, complementation and Cartesian products.

Example 5.1 (Examples of semi-algebraic sets and functions)**.** There is broad class of functions arising in optimization.

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.
- Finite sums and product of semi-algebraic functions.
- Composition of semi-algebraic functions.
- Sup/Inf type function, *e.g.*, $\sup\{g(u,v) : v \in C\}$ is semi-algebraic when g is a semi-algebraic function and C a semi-algebraic set.
- In matrix theory, all the following are semi-algebraic sets: cone of PSD matrices, Stiefel manifolds and constant rank matrices.
- The function $x \to \operatorname{dist}(x,S)^2$ is semi-algebraic whenever S is a nonempty semi-algebraic subset of $\mathbb{R}^d$.

Remark 5.1. The above results can be proven directly or via the fundamental Tarski-Seidenberg principle: The image of a semi-algebraic set $A \subset \mathbb{R}^{d+1}$ by the projection $\pi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ is semi-algebraic.

All these results and properties can be found in [1, 2, 3].

Let us now give some examples of semi-algebraic functions and other notions related to KL functions and their minimization through PALM.

Example 5.2 ($\|\cdot\|_0$ is semi-algebraic)**.** The sparsity measure (or the counting norm) of a vector x of $\mathbb{R}^d$ is defined by

$$\|x\|_0 := \text{ number of nonzero coordinates of } x.$$

For any given subset $I \subset \{1,\dots,d\}$, we denote by $|I|$ its cardinal and we define

$$J_i^I = \begin{cases} \{0\} & \text{if } i \in I, \\ \mathbb{R}\setminus\{0\} & \text{otherwise.} \end{cases}$$

The graph of $\|\cdot\|_0$ is given by a finite union of product sets:

$$\operatorname{graph}\|\cdot\|_0 = \bigcup_{I\subset\{1,\dots,d\}} \left(\prod_{i=1}^d J_i^I\right) \times \{d-|I|\},$$

it is thus a piecewise linear set, and in particular a semi-algebraic set. Therefore $\|\cdot\|_0$ is semi-algebraic. As a consequence the merit functions appearing in the various sparse NMF formulations we studied in Section 4 are semi-algebraic, hence KL.

Example 5.3 ($\|\cdot\|_p$ and KL functions)**.** Being given $p > 0$ the p norm is defined through

$$\|x\|_p = \left(\sum_{i=1}^{d} \|x_i\|^p \right)^{\frac{1}{p}}, \quad x \in \mathbb{R}^d.$$

Let us establish that $\|\cdot\|_p$ is semi-algebraic whenever p is rational, *i.e.*, $p = \frac{p_1}{p_2}$ where p_1 and p_2 are positive integers. From a general result concerning the composition of semi-algebraic functions we see that it suffices to establish that the function $s > 0 \to s^{\frac{p_1}{p_2}}$ is semi-algebraic. Its graph in $\mathbb{R}^2$ can be written as

$$\left\{ (s,t) \in \mathbb{R}^2_+ : \; t = s^{\frac{p_1}{p_2}} \right\} = \left\{ (s,t) \in \mathbb{R}^2 : \; t^{p_2} - s^{p_1} = 0 \right\} \cap \mathbb{R}^2_+.$$

This last set is semi-algebraic by definition.

When p is irrational $\|\cdot\|^p$ is not semi-algebraic, however for any semi-algebraic and lower semicontinuous functions H, f and any nonnegative real numbers α and λ the functions

$$\begin{aligned}
\Psi_1(x,y) &= f(x) + \lambda \|y\|_p + H(x,y) \\
\Psi_2(x,y) &= f(x) + \delta_{\|y\|_p \leq \alpha} + H(x,y) \\
\Psi_3(x,y) &= \delta_{\|x\|_p \leq \alpha} + \delta_{\|y\|_p \leq \alpha,\, y \geq 0} + H(x,y)
\end{aligned}$$

are KL functions (see, *e.g.*, [2] and references therein) with φ of the form $\varphi(s) = cs^{1-\theta}$ where c is positive and θ belongs to $(0,1]$.

Convex Functions and KL Property

Our developments on the convergence of PALM and its rate of convergence seem to be new even in the convex case. It is thus very important to realize that most convex functions encountered in finite dimensional applications satisfy the KL property. This may be due to the fact that they are semi-algebraic or subanalytic, but it can also come from more involved reasons involving o-minimal structures (see [2] for further details) or more down-to-earth properties like various growth conditions (see below). The reader which is wondering what a non KL convex function looks like can consult [15]. The convex counterexample provided in this work exhibit a wildly oscillatory collection of level sets, a phenomenon which seems highly unlikely to happen with functions modeling real world problems.

An interesting and rather specific feature of convex functions is that their desingularizing function φ can be explicitly computed from rather common and simple properties. Here are two important examples taken from [2].

Example 5.4 (Growth condition for convex functions)**.** Consider a proper, convex and lower semicontinuous function $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$. Assume that σ satisfies the following growth condition: There exist a neighborhood U of $\bar{x}$, $\eta > 0$, $c > 0$ and $r \geq 1$ such that

$$\forall x \in U \cap [\min \sigma < \sigma < \min \sigma + \eta], \; \sigma(x) \geq \sigma(\bar{x}) + c \cdot \operatorname{dist}(x, \operatorname{argmin} \sigma)^r,$$

where $\bar{x} \in \operatorname{argmin} \sigma \neq \emptyset$. Then σ satisfies the KL property at the point $\bar{x}$ for $\varphi(s) = r\, c^{-\frac{1}{r}}\, s^{\frac{1}{r}}$ on the set $U \cap [\min \sigma < \sigma < \min \sigma + \eta]$ (see, for more details, [15, 16]).

Example 5.5 (Uniform convexity)**.** Assume now that σ is uniformly convex *i.e.*, satisfies

$$\sigma(y) \geq \sigma(x) + \langle u, y - x \rangle + c \|y - x\|^p, \ p \geq 1$$

for all $x, y \in \mathbb{R}^d$ and $u \in \partial\sigma(x)$ (when $p = 2$ the function is called strongly convex). Then σ satisfies the Kurdyka-Łojasiewicz property on $\operatorname{dom} \sigma$ with $\varphi(s) = pc^{-\frac{1}{p}} s^{\frac{1}{p}}$.

References

[1] Attouch, H. and Bolte, J., On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Mathematical Programming* **116** (2009), 5–16.

[2] Attouch, H., Bolte, J., Redont, P. and Soubeyran, A., Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality, *Mathematics of Operations Research* **35** (2010), 438–457.

[3] Attouch, H., Bolte, J. and Svaiter, B. F., Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Mathematical Programming*, Ser. A **137** (2013), 91–129.

[4] Auslender, A., Méthodes numériques pour la décomposition et la minimisation de fonctions non différentiables, *Numerische Mathematik* **18** (1971), 213–223.

[5] Auslender, A., Optimisation – Méthodes numériques, Masson, Paris, 1976.

[6] Auslender, A., Asymptotic properties of the Fenchel dual functional and applications to decomposition problems, *Journal of Optimization Theory and Applications* **73** (1992), 427–449.

[7] Auslender, A., Teboulle, M. and Ben-Tiba, S., Coupling the Logarithmic-quadratic Proximal Method and the Block Nonlinear Gauss-Seidel Algorithm for Linearly Constrained Convex Minimization. In *Lecture Notes in Economics and Mathematical Systems*, Vol. 477, Eds. M. Thera and R. Tichastschke, pp. 35–47, (1998).

[8] Bauschke, H. H. and Combettes, P. L., Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer-Verlag, New York, 2011.

[9] Beck, A. and Teboulle, M., A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal of Imaging Sciences* **2** (2009), 183–202.

[10] Beck, A. and Tetruashvili, L., On the convergence of block coordinate descent type methods. Preprint (2011).

[11] Berry, M., Browne, M., Langville, A., Pauca, P. and Plemmons, R. J., Algorithms and applications for approximation nonnegative matrix factorization, *Computational Statistics and Data Analysis* **52** (2007), 155–173.

[12] Bertsekas, D. P. and Tsitsiklis, J. N., Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, New Jersey, 1989.

[13] Blum, M., Floyd, R. W., Pratt, V., Rivest, R. and Tarjan, R., Time bounds for selection, *Journal of Computer and System Sciences* **7** (1973), 448–461.

[14] Bolte J., Combettes, P.L., and Pesquet, J.-C, Alternating proximal algorithm for blind image recovery, in *Proceedings of the 17-th IEEE International Conference on Image Processing*, Hong-Kong, ICIP (2010), 1673–1676.

[15] Bolte, J., Daniilidis, A., Ley, O. and Mazet, L., Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity, *Transactions of the American Mathematical Society*, **362** (2010), 3319–3363.

[16] Bolte, J., Daniilidis, A. and Lewis, A., The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM Journal on Optimization* **17** (2006), 1205–1223.

[17] Bolte, J., Daniilidis, A., Lewis, A. and Shiota, M., Clarke subgradients of stratifiable functions, *SIAM Journal on Optimization* **18** (2007), 556–572.

[18] Cichocki, A., Zdunek, R., Phan, A. H. and Amari, S., Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation, John Wiley and Sons, 2009.

[19] Grippo, L. and Sciandrone, M., On the convergence of the block nonlinear Gauss-Seidel method under convex constraints, *Operations Research Letters* **26** (2000), 127–136.

[20] Heiler, M. and Schnorr, C., Learning sparse representations by non-negative matrix factorization and sequential cone programming, *Journal of Machine Learning Research* **7** (2006), 1385–1407.

[21] Hoyer, P. O., Non-negative matrix factorization with sparseness constraints, *Journal of Machine Learning Research* **5** (2004), 1457–1469.

[22] Kurdyka, K., On gradients of functions definable in o-minimal structures, *Annales de l'institut Fourier* **48** (1998), 769–783.

[23] Lee, D. D. and Seung, H. S., Learning the part of objects from nonnegative matrix factorization, *Nature* **401** (1999), 788–791.

[24] Lin C. J., Projected gradient methods for nonnegative matrix factorization, *Neural Computation* **19** (2007), 2756–2779.

[25] Łojasiewicz, S., Une propriété topologique des sous-ensembles analytiques réels, Les Équations aux Dérivées Partielles. Éditions du centre National de la Recherche Scientifique, Paris, 87–89, (1963).

[26] Luss, R. and Teboulle, M., Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint, *SIAM Review* **55** (2013), 65–98.

[27] Mordukhovich, B., Variational Analysis and Generalized Differentiation. I. Basic Theory, Grundlehren der Mathematischen Wissenschaften, 330, Springer-Verlag, Berlin, 2006.

[28] Nesterov. Y., Introductory Lectures on Convex Optimization, Kluwer, Boston, 2004.

[29] Ortega, J. M. and Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New-York, 1970.

[30] Palomar, D. P. and Eldar, Y. (Eds.), Convex Optimization in Signal Processing and Communications, Cambridge University Press, UK, 2010.

[31] Polak, E., Sargent, R. W. H. and Sebastian, D. J., On the convergence of sequential minimization algorithms, *Journal of Optimization Theory and Applications* **14** (1974), 439–442.

[32] Powell, M. J. D., On search directions for minimization algorithms, *Mathematical Programming* **4** (1973), 193–201.

[33] Rockafellar, R. T. and Wets, R., Variational Analysis, Grundlehren der Mathematischen Wissenschaften, 317, Springer, 1998.

[34] Sra, S., Nowozin, S. and Wright, S. J. (Eds.), Optimization for Machine Learning, The MIT Press, Cambridge, 2011.

[35] Tseng, P., Convergence of a block coordinate descent method for nondifferentiable minimization, *Journal of Optimization Theory and Applications* **109** (2001), 475–494.

[36] Zangwill, W. I., Nonlinear Programming: A Unified Approach, Prentice Hall, Englewood Cliffs, 1969.