SPLITTING METHODS WITH VARIABLE METRIC FOR KŁ FUNCTIONS AND GENERAL CONVERGENCE RATES

PIERRE FRANKEL, GUILLAUME GARRIGOS, JUAN PEYPOUQUET

Abstract. We study the convergence of general abstract descent methods applied to a lower semicontinuous nonconvex function f that satisfies the Kurdyka-Łojasiewicz inequality in a Hilbert space. We prove that any precompact sequence converges to a critical point of f and obtain new convergence rates both for the values and the iterates. The analysis covers alternating versions of the forward-backward method with variable metric and relative errors. As an example, a nonsmooth and nonconvex version of the Levenberg-Marquardt algorithm is detailled.

Key words: Nonconvex and nonsmooth optimization ; Kurdyka-Łojasiewicz inequality ; Descent methods ; Convergence rates ; Variable metric ; Gauss-Seidel method ; Newton-like method

AMS subject classification. 49M37, 65K10, 90C26, 90C30

The second and third authors are partly supported by Conicyt Anillo Project ACT-1106, ECOS-Conicyt Project C13E03 and Millenium Nucleus ICM/FIC P10-024F. The third author is also partly supported by FONDECYT Grant 1140829 and Basal Project CMM Universidad de Chile.

1. Introduction

In this paper we present a class of numerical methods to find critical points for a class of nonsmooth and nonconvex functions defined on a Hilbert space. Our analysis relies on the Kurdyka-Łojasiewicz (KŁ) inequality, initially formulated by Łojasiewicz for analytic functions in finite dimension [1], and later extended to nonsmooth functions in more general spaces [2, 3, 4, 5]. Gradient-like systems governed by potentials satisfying this KŁ inequality enjoy good asymptotic properties: under a compactness assumption, the corresponding trajectories have finite length and converge strongly to equilibria or critical points. These ideas were used in [6] to study nonlinear first-order evolution equations (see also [7, 8]). Second-order systems were considered in [9, 10] and a Schrödinger equation in [11].

The convergence analysis of algorithms in this context is more recent. See [12] for gradient-related methods, [13, 14, 15] for the proximal point algorithm and [16] for a nonsmooth subgradient-oriented descent method. The celebrated Forward-Backward algorithm, a splitting method exploiting the nonsmooth/smooth structure of the objective function, has been studied in [17], and extended in [18] to take in account a variable metric. Another splitting approach comes from Gauss-Seidel-like methods, which apply to functions with separated variables, and consist in doing a descent method relatively to each (block of) variables alternatively. See [19, 20] for a proximal alternating

P. Frankel (p.frankel30@orange.fr) & G. Garrigos (guillaume.garrigos@gmail.com)
Institut de Mathématiques et Modélisation de Montpellier, UMR 5149 CNRS.
Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier cedex 5, France.
G. Garrigos & J. Peypouquet (juan.peypouquet@usm.cl)
Departamento de Matemática & AM2V.
Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile.

method, and [17] for a variable-metric version. Recent papers [20, 21, 22] propose to combine these two splitting approaches in order to exploit both the smooth/nonsmooth character and the separated structure of the function.

Most of the algorithms studied in the aforementioned papers share the same asymptotic behavior: under a compactness assumption, the sequences generated converge strongly to critical points, and the affine interpolations have finite length. This is not surprising since the algorithms described in [13, 15, 14, 17, 19, 21] together with the ones of [18, 20] (without extrapolation step) fall into the general convergence result for abstract descent methods of Attouch, Bolte and Svaiter [17]. Besides, these methods essentially share the same hypotheses on the parameters with the abstract method of [17] : the step sizes (resp. the eigenvalues of the matrices underlying the metric) are required to remain in a compact subinterval of the positive numbers. Moreover they have little flexibility regarding the presence of computational errors. To our knowledge vanishing step sizes (resp. unbounded eigenvalues) or sufficiently general errors have never been treated in the KŁ context.

Another interesting aspect is that the convergence rate of several of these methods are essentially the same, and depend on the KŁ inequality rather than the nature of the algorithm. Therefore, it seems reasonable to consider the existence of an abstract convergence rate result for general descent methods.

We present now the structure of the paper and underline its main contributions: in Section 2 we recall some definitions, well-known facts, and set the notation. Section 3 contains the main *theoretical* results of the paper. More precisely, in Subsection 3.1, we present an abstract inexact descent method, which is inspired by [17] but extending their setting in order to account for additive computational errors and more versatility in the choice of the parameters. The strong convergence of the iterates with a *finite-length* condition, and a *capture* property are proved under certain hypotheses. Since the proofs are very close to those of [17], most arguments are given in Appendix A.1. Then, in Subsection 3.2 we prove new and interesting general convergence rates. They are similar to the ones obtained in [13, 19, 20, 21, 22]. Surprisingly, an explicit form of the algorithm terminates in a finite number of iterations in several cases. A link with convergence rates for some continuous-time dynamical systems is also given. Sections 4 and 5 contain the main *practical* contributions. In Section 4, we present a particular instance of the model, which provides further insight into a large class of known methods and present some innovative variants. More exactly, we revisit the *Alternating Forward-Backward* methods, already considered in [21, 22, 23], but allowing inexact computation of the iterates and a dynamic choice of metric. This setting includes also the *generalized Levenberg-Marquardt algorithm*, a Newton-like method adapted for nonconvex and nonsmooth functions. In Section 5, we briefly describe an instance of this algorithm to produce a new method for the sparse and low-rank matrix decomposition. Finally, some perspectives are discussed in Section 6.

2. Preliminaries

Throughout this paper H is a real Hilbert space with norm $\|\cdot\|$ and scalar product $\langle\cdot,\cdot\rangle$. We write $x^k \longrightarrow x$, or $x^k \xrightarrow{w} x$, if x^k converges strongly or weakly to x, respectively, as $k \to +\infty$. The *domain* of $f : H \longrightarrow \mathbb{R} \cup \{+\infty\}$ is $\operatorname{dom} f = \{x : f(x) < +\infty\}$. A sequence x^k *f-converges to x* (we write $x^k \xrightarrow{f} x$) if $x^k \longrightarrow x$ and $f(x^k) \longrightarrow f(x)$. We say that a sequence is *precompact* (resp. *f-precompact*) if it has at least one convergent (resp. f-convergent) subsequence.

2.1. **Subdifferential and critical points.** Let $f : H \to \mathbb{R} \cup \{+\infty\}$. The *Fréchet subdifferential* of f at $x \in \operatorname{dom} f$ is the set $\partial_F f(x)$ of those elements $p \in H$ such that

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

For $x \notin \operatorname{dom} f$, we set $\partial_F f(x) := \emptyset$. The *(limiting Fréchet) subdifferential* of f at $x \in \operatorname{dom} f$ is the set $\partial f(x)$ of elements $p \in H$ for which there exists sequences $(x^k)_{k\in\mathbb{N}}$ and $(p^k)_{k\in\mathbb{N}}$ in H such that $x^k \xrightarrow{f} x$, $p^k \xrightarrow{w} p$, and $p^k \in \partial_F f(x^k)$. As before, $\partial f(x) := \emptyset$ for $x \notin \operatorname{dom} f$ and its domain is $\operatorname{dom} \partial f := \{x \in H : \partial f(x) \neq \emptyset\}$. This subdifferential satisfies the following chain rule : let g_1, g_2 and h be extended real valued functions on H_1, H_2 and $H_1 \times H_2$ respectively. If h is continuously differentiable in a neighbourhood of $(x_1, x_2) \in \operatorname{dom} g_1 \times \operatorname{dom} g_2$, the subdifferential of $f(x_1, x_2) := g_1(x_1) + g_2(x_2) + h(x_1, x_2)$ at (x_1, x_2) is

$$\partial f(x_1, x_2) = \Big(\partial g_1(x_1) + \{\nabla_1 h(x_1, x_2)\}\,,\, \partial g_2(x_2) + \{\nabla_2 h(x_1, x_2)\}\Big). \tag{1}$$

We say that $x \in H$ is a *critical point* if $0 \in \partial f(x)$. The *lazy slope* of f at x is $\|\partial f(x)\|_- := \inf_{p \in \partial f(x)} \|p\|$ if $x \in \operatorname{dom} \partial f$, and $+\infty$ otherwise. This definition gives the following result:

Lemma 1. *If* $x^k \xrightarrow{f} x$ *and* $\liminf_{n\to+\infty} \|\partial f(x^k)\|_- = 0$, *then* $0 \in \partial f(x)$.

2.2. **The Kurdyka-Łojasiewicz property.** Let $\eta \in]0, +\infty]$ and let $\varphi : [0, \eta[\longrightarrow [0, +\infty[$ be a continuous concave function such that $\varphi(0) = 0$ and φ is continuously differentiable on $]0, \eta[$ with $\varphi'(t) > 0$ for all $t \in]0, \eta[$. A proper lower-semicontinuous function $f : H \to \mathbb{R} \cup \{+\infty\}$ has the *Kurdyka-Łojasiewicz property* at a point $x^* \in \operatorname{dom} \partial f$ with *desingularizing function* φ if there exists $\delta > 0$ such that the *Kurdyka-Łojasiewicz inequality*

$$\varphi'(f(x) - f(x^*))\|\partial f(x)\|_- \geq 1 \tag{2}$$

holds for all x in the *strict local upper level* set

$$\Gamma_\eta(x^*, \delta) = \{\, x \in H \ :\ \|x - x^*\| < \delta \text{ and } f(x^*) < f(x) < f(x^*) + \eta \,\}. \tag{3}$$

A proper lower-semicontinuous function having the Kurdyka-Łojasiewicz property at each point of $\operatorname{dom} \partial f$ is a *KŁ function*. When f is continuously differentiable, (2) becomes $\|\nabla(\varphi \circ f)\| \geq 1$. This means that the more f is flat around its critical points, the more φ has to be steep around 0, whence the term "desingularizing". The KŁ property reveals the possibility to reparameterize the values of f in order to avoid flatness around the critical points. We shall see in Subsection 3.2 that the growth of φ has a direct impact on the convergence rate of optimization algorithms.

Semi-algebraic and bounded sub-analytic functions in finite dimension satisfy a KŁ inequality ([3, 4, 5]), as well as some, but not all, convex functions (see [14] for details and a counterexample). See [24, 25, 26], and the references therein, for more information in the general context of *o-minimal* functions. See [27, 28] for characterizations in infinite-dimensional Hilbert spaces.

2.3. **Proximal operator in a given metric.** Let $\mathcal{S}_{++}(H)$ denote the space of bounded, uniformly elliptic and self-adjoint operators on H. Each $A \in \mathcal{S}_{++}(H)$ induces a metric on H by the inner product $\langle x, y\rangle_A := \langle Ax, y\rangle$, and the norm $\|x\|_A := \sqrt{\langle x, x\rangle_A}$. We also set $\alpha(A)$ as the infimum of the spectral values of A, satisfying $\|x\|_A^2 \geq \alpha(A)\|x\|^2$

for all $x \in H$. Let $f : H \to \mathbb{R} \cup \{+\infty\}$, the *proximal operator of f in the metric induced by A* is the set-valued mapping $\operatorname{prox}_f^A : H \rightrightarrows H$, defined as

$$\operatorname{prox}_f^A(x) := \operatorname*{argmin}_{y \in H} \left\{ f(y) + \frac{1}{2}\|y - x\|_A^2 \right\}. \tag{4}$$

Observe that $\operatorname{prox}_f^A(x) \neq \emptyset$ if f is weakly lower-semicontinuous and bounded from below (see [29, Theorem 3.2.5]), which holds in many relevant applications. If f is the indicator function of a set, then $\operatorname{prox}_f^A(x)$ is the *nearest point mapping* relatively to the metric induced by A.

3. Convergence of an abstract inexact descent method

Throughout this section, $f : H \to \mathbb{R} \cup \{+\infty\}$ is a proper function that is lower-semicontinuous for the strong topology. We shall adopt the notation given in Subsection 2.2 concerning the KŁ property, whenever it is invoked. We consider a sequence $(x^k)_{k\in\mathbb{N}}$, computed by means of an abstract algorithm satisfying the following hypotheses:

$\mathbf{H}_1$ (*Sufficient decrease*): For each $k \in \mathbb{N}$, for some $a_k > 0$,

$$f(x^{k+1}) + a_k\|x^{k+1} - x^k\|^2 \leq f(x^k).$$

$\mathbf{H}_2$ (*Relative error*): For each $k \in \mathbb{N}$, for some $b_{k+1} > 0$ and $\varepsilon_{k+1} \geq 0$,

$$b_{k+1}\|\partial f(x^{k+1})\|_- \leq \|x^{k+1} - x^k\| + \varepsilon_{k+1}.$$

$\mathbf{H}_3$ (*Parameters*): The sequences $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ and $(\varepsilon_k)_{k\in\mathbb{N}}$ satisfy:

(i) $a_k \geq \underline{a} > 0$ for all $k \geq 0$.
(ii) $(b_k)_{k\in\mathbb{N}} \notin l^1$;
(iii) $\sup_{k\in\mathbb{N}^*} \frac{1}{a_k b_k} < +\infty$;
(iv) $(\epsilon_k)_{k\in\mathbb{N}} \in l^1$.

In Section 4, we complement this axiomatic description of descent methods by providing a large class of implementable algorithms that produce sequences verifying hypotheses $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$. A simple example is:

Example 1. If f is differentiable, a gradient-related method (see [30]) is an algorithms where each iteration has the form $x^{k+1} = x^k + \lambda_k d^k$, where $\lambda_k > 0$ and d^k agrees with the steepest descent direction $-\nabla f(x^k)$ in the sense that $\langle d^k, \nabla f(x^k)\rangle + C\|d^k\|^2 \leq 0$ and $\|\nabla f(x^k) + d^k\| \leq C\|d^k\| + e_k$, with $C > 0$ and $\lim_{k\to\infty} e_k = 0$. If ∇f is Lipschitz-continuous, it is easy to find conditions on the sequence (λ_k) to verify hypotheses $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$.

3.1. Capture, convergence, and finite length of the trajectories.

Sequences generated by the procedure described above converge strongly to critical points of f and the piecewise linear curve obtained by interpolation has finite length. More precisely, we have:

Theorem 1. *Let $f : H \to \mathbb{R} \cup \{+\infty\}$ be a KŁ function and let $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ hold. If the sequence $(x^k)_{k\in\mathbb{N}}$ is f-precompact, then it f-converges to a critical point of f and $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$.*

It is possible in Theorem 1 to drop the f-precompactness assumption and obtain a *capture result*, near a global minimum of f. To simplify the notation, for $x^* \in H$, $\eta \in]0, +\infty]$ and $\delta > 0$, define the *relaxed local upper level set* by

$$\underline{\Gamma}_\eta(x^*, \delta) = \{\, x \in H \;:\; \|x - x^*\| < \delta \text{ and } f(x^*) \leq f(x) < f(x^*) + \eta \,\}. \tag{5}$$

We have the following:

Theorem 2. *Let $f : H \longrightarrow \mathbb{R} \cup \{+\infty\}$ have the KŁ property in a global minimum x^* of f. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence satisfying $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ with $\epsilon_k \equiv 0$. Then, there exist $\gamma > 0$ and $\eta > 0$ such that if $x^0 \in \underline{\Gamma}_\eta(x^*, \gamma)$, then $(x^k)_{k\in\mathbb{N}}$ f-converges to a global minimum $\overline{x}$ of f, with $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$.*

As mentioned in [17], Theorem 2 admits a more general formulation, for instance, if x^* is a local minimum of f where a growth assumption is locally satisfied (see [17, Remark 2.11]).

The proofs of Theorems 1 and 2 follow the arguments in [17, Subsection 2.3], adapted to the presence of errors and the variability of the parameters. They are given in Appendix A.1 for the reader's convenience.

3.2. **Rates of Convergence.** We assume that $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ hold, and for simplicity and precision, we restrict ourselves to the case where $\varepsilon_k \equiv 0$. Suppose that x^k f-converges to a point x^* where f has the KŁ property. We study three types of convergence rate results, depending on the nature of the desingularizing function φ:

1. Theorem 3 establishes the relationship between the distance to the limit $\|x^k - x^*\|$ and the gap $f(x^k) - f(x^*)$, for a generic desingularizing function. It is similar to the result in [14] for the proximal method in the convex case.
2. Theorem 4 gives explicit convergence rates in terms of the parameters − both for the distance and the gap − when the desingularizing function is of the form $\varphi(t) = \frac{C}{\theta} t^\theta$ with $C > 0$ and $\theta \in]0, 1]$. Several results obtained in the literature for various methods are recovered.
3. Finally, Theorem 5 provides convergence rates when $\mathbf{H}_2$ is replaced by a slightly different hypothesis that holds for certain explicit schemes, namely gradient-related methods. This result is valid for a generic desingularizing function φ. However, when φ is of the form $\varphi(t) = \frac{C}{\theta} t^\theta$ ($C > 0$, $\theta \in]0, 1]$) the prediction is considerably better than the one provided by Theorem 4.

3.2.1. *Distance to the limit in terms of the gap.*

Theorem 3. *Set $\tilde{\varphi}(t) := \max\{\varphi(t), \sqrt{t}\}$. Then $\|x^* - x^k\| = O\left(\tilde{\varphi}(f(x^{k-1}) - f(x^*))\right)$.*

Proof. By assumption, $x^k \xrightarrow{f} x^*$ and f satisfies the KŁ inequality on some $\Gamma_\eta(x^*, \delta)$. Let $r_k := f(x^k) - f(x^*) \geq 0$. We may suppose that $r_k > 0$ for all $k \in \mathbb{N}$ because otherwise the algorithm terminates in a finite number of steps. For K large enough, we have $x^k \in \Gamma_\eta(x^*, \delta)$ for all $k \geq K$. Lemma 2, gives

$$2\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + M[\varphi(r_k) - \varphi(r_{k+1})]$$

for all $k \geq K$. Summing this inequality for $k = K, \ldots, N$, we obtain

$$\sum_{k=K}^{N} \|x^{k+1} - x^k\| \leq \|x^K - x^{K-1}\| + M\varphi(r_K).$$

Using the triangle inequality and passing to the limit, we get

$$\|x^*-x^K\| \leq \sum_{k=K}^{\infty} \|x^{k+1}-x^k\| \leq \|x^K-x^{K-1}\|+M\varphi(r_K) \leq \frac{\sqrt{f(x^{K-1})-f(x^K)}}{\sqrt{a_K}}+M\varphi(r_K)$$

by $\mathbf{H}_1$. Then, using $\mathbf{H}_0$, along with the fact that $f(x^K) \geq f(x^*)$ and that (r_k) is decreasing, we deduce that $\|x^* - x^K\| \leq \frac{1}{\sqrt{\underline{a}}}\sqrt{r_{K-1}} + M\varphi(r_{K-1})$, which finally gives $\|x^* - x^K\| \leq \max\left\{\frac{1}{\sqrt{\underline{a}}}, M\right\}\tilde{\varphi}(r_{K-1})$. □

3.2.2. *Explicit rates when $\varphi(t) = \frac{C}{\theta}t^\theta$ with $C > 0$ and $\theta \in]0,1]$.* Theorem 4 below is qualitatively analogous to the results in [13, 15, 19, 20, 21, 22] : convergence in a finite number of steps if $\theta = 1$, exponential convergence if $\theta \in [\frac{1}{2}, 1[$ and polynomial convergence if $\theta \in]0, \frac{1}{2}[$. In the general convex case, finite-time termination of the proximal point algorithm was already proved in [31] and [32] (see also [33]).

Theorem 4. *Assume $\varphi(t) = \frac{C}{\theta}t^\theta$ for some $C > 0$, $\theta \in]0,1]$.*

i) If $\theta = 1$ and $\inf_{k\in\mathbb{N}} a_k b_{k+1}^2 > 0$,[1] then x^k converges in finite time.

ii) If $\theta \in [\frac{1}{2}, 1[$, $\sup_{k\in\mathbb{N}} b_k < +\infty$ and $\inf_{k\in\mathbb{N}} a_k b_{k+1} > 0$,[1] there exist $c > 0$ and $k_0 \in \mathbb{N}$ such that:

1. $f(x^k) - f(x^*) = O\left(\exp\left(-c\sum_{n=k_0}^{k-1} b_{n+1}\right)\right)$, *and*

2. $\|x^* - x^k\| = O\left(\exp\left(-\frac{c}{2}\sum_{n=k_0}^{k-2} b_{n+1}\right)\right)$.

iii) If $\theta \in]0, \frac{1}{2}[$, $\sup_{k\in\mathbb{N}} b_k < +\infty$ and $\inf_{k\in\mathbb{N}} a_k b_{k+1} > 0$,[1] there is $k_0 \in \mathbb{N}$ such that:

1. $f(x^k) - f(x^*) = O\left(\left(\sum_{n=k_0}^{k-1} b_{n+1}\right)^{\frac{-1}{1-2\theta}}\right)$, *and*

2. $\|x^* - x^k\| = O\left(\left(\sum_{n=k_0}^{k-2} b_{n+1}\right)^{\frac{-\theta}{1-2\theta}}\right)$.

Proof. We can suppose that $r_k > 0$ for all $k \in \mathbb{N}$, because otherwise the algorithm terminates in a finite number of steps. Since x^k converges to x^*, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$ we have $x^k \in \Gamma_\eta(x^*, \delta)$ where the KŁ inequality holds. Using successively $\mathbf{H}_1$, $\mathbf{H}_2$ and the KŁ inequality we obtain

$$\varphi'^2(r_{k+1})(r_k - r_{k+1}) \geq \varphi'^2(r_{k+1})a_k b_{k+1}^2 \|\partial f(x^{k+1})\|_-^2 \geq a_k b_{k+1}^2 \tag{6}$$

for each $k \geq k_0$. Let us now consider different cases for θ:
Case $\theta = 1$: If $r_k > 0$ for all $k \in \mathbb{N}$, then $C^2(r_k - r_{k+1}) \geq a_k b_{k+1}^2 \geq \inf_{k\in\mathbb{N}} a_k b_{k+1}^2 > 0$ for all $k \geq k_0$. Since r_k converges, we must have $\inf_{k\in\mathbb{N}} a_k b_{k+1}^2 = 0$, which is a contradiction. Therefore, there exists some $k \in \mathbb{N}$ such that $r_k = 0$, and the algorithm terminates in a finite number of steps.

[1] A simple sufficient − yet not necessary − condition for $\inf_{k\in\mathbb{N}} a_k b_{k+1}^2 > 0$ and $\inf_{k\in\mathbb{N}} a_k b_{k+1} > 0$ is that $\inf_{k\in\mathbb{N}} b_k > 0$.

Case $\theta \in]0,1[$: Write $\bar{b} := \sup_{k\in\mathbb{N}} b_k$, $m := \inf_{k\in\mathbb{N}} a_k b_{k+1}$ and $c = \frac{m}{C^2(1+\bar{b})}$ and, for each $k \in \mathbb{N}$, $\beta_k := \frac{b_k m}{C^2}$. For each $k \geq k_0$, inequality (6) gives

$$(r_k - r_{k+1}) \geq \frac{a_k b_{k+1}^2 r_{k+1}^{2-2\theta}}{C^2} \geq \beta_{k+1} r_{k+1}^{2-2\theta}. \tag{7}$$

Subcase $\theta \in [\frac{1}{2}, 1[$: Since $r_k \to 0$ and $0 < 2 - 2\theta \leq 1$, we may assume, by enlarging k_0 if necessary, that $r_{k+1}^{2-2\theta} \geq r_{k+1}$ for all $k \geq k_0$. Inequality (7) implies $(r_k - r_{k+1}) \geq \beta_{k+1} r_{k+1}$ or, equivalently, $r_{k+1} \leq r_k \left(\frac{1}{1+\beta_{k+1}}\right)$ for all $k \geq k_0$. By induction, we obtain

$$r_{k+1} \leq r_{k_0} \left(\prod_{n=k_0}^{k} \frac{1}{1+\beta_{n+1}}\right) = r_{k_0} \exp\left(\sum_{n=k_0}^{k} \ln\left(\frac{1}{1+\beta_{n+1}}\right)\right)$$

for all $k \geq k_0$. But $\ln\left(\frac{1}{1+\beta_{n+1}}\right) \leq \frac{-\beta_{n+1}}{1+\beta_{n+1}} \leq \frac{-1}{1+\bar{b}}\beta_{n+1}$, and so

$$r_{k+1} \leq r_{k_0} \exp\left\{\sum_{n=k_0}^{k} \left(\frac{-1}{1+\bar{b}}\beta_{n+1}\right)\right\} = r_{k_0} \exp\left(-c \sum_{n=k_0}^{k} b_{n+1}\right).$$

The second part follows from Theorem 3.

Subcase $\theta \in]0, \frac{1}{2}[$: Recall from inequality (7) that $r_{k+1}^{2\theta-2}(r_k - r_{k+1}) \geq \beta_{k+1}$. Set $\phi(t) := \frac{C}{1-2\theta} t^{2\theta-1}$. Then $\phi'(t) = -Ct^{2\theta-2}$, and

$$\phi(r_{k+1}) - \phi(r_k) = \int_{r_k}^{r_{k+1}} \phi'(t)\, dt = C \int_{r_{k+1}}^{r_k} t^{2\theta-2}\, dt \geq C(r_k - r_{k+1}) r_k^{2\theta-2}.$$

On the one hand, if we suppose that $r_{k+1}^{2\theta-2} \leq 2r_k^{2\theta-2}$, then

$$\phi(r_{k+1}) - \phi(r_k) \geq \frac{C}{2}(r_k - r_{k+1}) r_{k+1}^{2\theta-2} \geq \frac{C}{2}\beta_{k+1}.$$

On the other hand, suppose that $r_{k+1}^{2\theta-2} > 2r_k^{2\theta-2}$. Since $2\theta - 2 < 2\theta - 1 < 0$, we have $\frac{2\theta-1}{2\theta-2} > 0$. Thus $r_{k+1}^{2\theta-1} > q r_k^{2\theta-1}$, where $q := 2^{\frac{2\theta-1}{2\theta-2}} > 1$. Therefore,

$$\phi(r_{k+1}) - \phi(r_k) = \frac{C}{1-2\theta}(r_{k+1}^{2\theta-1} - r_k^{2\theta-1}) > \frac{C}{1-2\theta}(q-1) r_k^{2\theta-1} \geq C',$$

with $C' := \frac{C}{1-2\theta}(q-1) r_{k_0}^{2\theta-1} > 0$. Since $\beta_{k+1} \leq \frac{\bar{b}m}{C^2}$, we can write

$$\phi(r_{k+1}) - \phi(r_k) \geq \frac{C'C^2}{\bar{b}m}\beta_{k+1}.$$

Setting $c := \min\{\frac{C}{2}, \frac{C'C^2}{\bar{b}m}\} > 0$ we can write $\phi(r_{k+1}) - \phi(r_k) \geq c\beta_{k+1}$ for all $k \geq k_0$. This implies

$$\phi(r_{k+1}) \geq \phi(r_{k+1}) - \phi(r_{k_0}) = \sum_{n=k_0}^{k} \phi(r_{n+1}) - \phi(r_n) \geq c \sum_{n=k_0}^{k} \beta_{n+1},$$

which is precisely $r_{k+1} \leq D\left(\sum_{n=k_0}^{k} b_{n+1}\right)^{\frac{-1}{1-2\theta}}$ with $D = \left(\frac{cm(1-2\theta)}{C^3}\right)^{\frac{-1}{1-2\theta}}$. As before, Theorem 3 gives the second part. $\square$

3.2.3. *Sharper results for gradient-related methods.* Convergence rates for the continuous-time gradient system

$$-\dot{x}(t) = \nabla f(x(t)), \tag{8}$$

where f is some integral functional, are given in [34]. For any φ, [34, Theorem 2.7] states that

(1) $f(x^k) - f(x^*) = O\left(\Phi^{-1}(t-\hat{t})\right)$, and
(2) $\|x^* - x^k\|_{L^2(\Omega)} = O\left(\varphi \circ \Phi^{-1}(t-\hat{t})\right)$,

where Φ is any primitive of $-(\varphi')^2$. If the desingularizing function φ has the form $\varphi(t) = \frac{Ct^\theta}{\theta}$, we recover (see [34, Remark 2.8]) convergence in finite time if $\theta \in]\frac{1}{2}, 1]$, exponential convergence if $\theta = \frac{1}{2}$, and polynomial convergence if $\theta \in]0, \frac{1}{2}[$. The same conclusion was established in [5, Theorem 4.7] for a nonsmooth version of (8) when f is any subanalytic function in $\mathbb{R}^N$. This prediction is better than the one given by Theorem 4 above, as well as the results in [13, 15, 19, 20, 21, 22] since it guarantees convergence in finite time for $\theta > \frac{1}{2}$. We shall prove that for certain algorithms including gradient-related methods, this better estimation remains true. To this end, consider the following variant of hypothesis $\mathbf{H}_2$:

$\underline{\mathbf{H}_2'\ (\textit{Relative error})}$: For each $k \in \mathbb{N}$, $b_{k+1}\|\partial f(x^k)\|_- \leq \|x^{k+1} - x^k\|$.

Theorem 5. *Suppose condition $\mathbf{H}_2'$ is satisfied instead of $\mathbf{H}_2$ and assume $m := \inf_{k\in\mathbb{N}} a_k b_{k+1} > 0$. Let $\Phi :]0, \eta[\to \mathbb{R}$ be any primitive of $-(\varphi')^2$.*

i) If $\lim_{t\to 0} \Phi(t) \in \mathbb{R}$, then the algorithm converges in a finite number of steps.

ii) If $\lim_{t\to 0} \Phi(t) = +\infty$, then there exists $k_0 \in \mathbb{N}$ such that:

1. $f(x^k) - f(x^*) = O\left(\Phi^{-1}\left(m \sum_{n=k_0}^{k-1} b_{n+1}\right)\right)$, *and*
2. $\|x^* - x^k\| = O\left(\varphi \circ \Phi^{-1}\left(m \sum_{n=k_0}^{k-1} b_{n+1}\right)\right)$.

Proof. The following proof is inspired by the one of [34] in the continuous case. First, if $r_k > 0$ for all $k \in \mathbb{N}$, we claim that there is $k_0 \in \mathbb{N}$ such that

$$\Phi(r_{k+1}) \geq \Phi(r_{k_0}) + m \sum_{n=k_0}^{k} b_{n+1}. \tag{9}$$

To see this, let k_0 be large enough to have $x^k \in \Gamma_\eta(x^*, \delta)$ where the KŁ inequality holds for all $k \geq k_0$. We apply successively $\mathbf{H}_1$, $\mathbf{H}_2'$, the KŁ inequality and $\mathbf{H}_3$ to obtain

$$\varphi'(r_k)^2(r_k - r_{k+1}) \geq \varphi'(r_k)^2 a_k b_{k+1}^2 \|\partial f(x^k)\|_-^2 \geq a_k b_{k+1}^2 \geq b_{k+1} m.$$

Let Φ be a primitive of $-(\varphi')^2$. Then

$$\Phi(r_{k+1}) - \Phi(r_k) = \int_{r_{k+1}}^{r_k} \varphi'(t)^2\, dt \geq (r_k - r_{k+1})\varphi'(r_k)^2 \geq b_{k+1} m$$

because φ' is decreasing. Therefore,

$$\Phi(r_{k+1}) - \Phi(r_{k_0}) = \sum_{n=k_0}^{k} \Phi(r_{n+1}) - \Phi(r_n) \geq m \sum_{n=k_0}^{k} b_{n+1}$$

as claimed. Now let us analyze the two cases:

For i), if $r_k > 0$ for all $k \in \mathbb{N}$, then (9) implies $\lim_{k\to+\infty} \Phi(r_{k+1}) = +\infty$ which contradicts the fact that $\lim_{t\to 0} \Phi(t) \in \mathbb{R}$. Hence, $r_k = 0$ for some $k \in \mathbb{N}$.

For ii), we may suppose that $r_k > 0$ for all $k \in \mathbb{N}$ (otherwise the algorithm stops in a finite number of steps) and so (9) holds for all $k \in \mathbb{N}$. Since $\lim_{k\to+\infty} \Phi(r_k) = +\infty$, we can take k_0 large enough to have $\Phi(r_{k_0}) > 0$. Whence $\Phi(r_{k+1}) \geq m \sum_{n=k_0}^{k} b_{n+1}$. Since $(b_n) \notin \ell^1$, for all sufficiently large k, $m \sum_{n=k_0}^{k} b_{n+1}$ is in the domain of Φ^{-1} and we obtain the first estimation, namely:

$$r_{k+1} \leq \Phi^{-1}\left(m \sum_{n=k_0}^{k} b_{n+1}\right). \tag{10}$$

For the second one, since φ is concave and differentiable, we have

$$\varphi(r_k) - \varphi(r_{k+1}) \geq \varphi'(r_k)(r_k - r_{k+1}) \geq \varphi'(r_k) a_n \|x^{k+1} - x^k\|^2,$$

by $\mathbf{H}_1$. The KŁ property and $\mathbf{H}_2'$ then give

$$\varphi(r_k) - \varphi(r_{k+1}) \geq m\|x^{k+1} - x^k\|,$$

which in turn yields

$$\|x^* - x^k\| \leq \frac{1}{m} \sum_{n=k}^{\infty} [\varphi(r_n) - \varphi(r_{n+1})] \leq \frac{1}{m}\varphi(r_k).$$

We conclude by using (10). □

4. Descent methods with errors and variable metric

As stressed in [17], the abstract scheme developed in Section 3 covers, among others, the gradient-related methods (a wide variety of schemes based on the gradient method sketched in [35]), the proximal algorithm (introduced in [36] and further developed in [37, 31]), and the forward-backward algorithm (a combination of the preceding, see [38, 39]). This last one is a splitting method, used to solve structured optimization problems with the following form

$$\underset{x\in H}{\text{minimize}}\ f(x) = g(x) + h(x), \tag{11}$$

where g is a nonsmooth proper l.s.c function and h is differentiable with a L Lipschitz gradient. It has been studied in the nonsmooth nonconvex setting in [17] and the algorithm was stated as follows: start with $x^0 \in H$, consider $(\lambda_k) \subset [\underline{\lambda}, \bar{\lambda}]$ with $0 < \underline{\lambda} \leq \bar{\lambda} < \frac{1}{L}$ and $\forall k \in \mathbb{N}$

$$x^{k+1} \in \operatorname{prox}_{\lambda_k g}\left(x^k - \lambda_k \nabla h(x^k)\right). \tag{12}$$

It satisfies $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ (see [17, Theorem 5.1]) and falls into the setting of Theorem 1. We shall extend this class of algorithms in different directions:

- Alternative choice of metric for the ambient space, which may vary at each step (see [40, 41] and the references therein). Considering metrics induced by a sequence $(A_k) \subset \mathcal{S}_{++}(H)$, the forward-backward method becomes

$$x^{k+1} \in \operatorname{prox}_g^{A_k}\left(x^k - A_k^{-1}\nabla h(x^k)\right) \tag{13}$$

(recall Subsection 2.3). Indeed, (13) can be rewritten as

$$x^{k+1} \in \operatorname*{argmin}_{y\in H} \; g(y) + h(x^k) + \langle y - x^k, \nabla h(x^k)\rangle + \frac{1}{2}\langle y - x^k, A_k(y - x^k)\rangle. \tag{14}$$

At each step, an approximation of f, replacing its smooth part h by a quadratic model, is minimized. See [18] for a similar algorithm called Variable Metric Forward-Backward, and [16] for an approach considering more general models. Note that when $A_k = \frac{1}{\lambda_k} id_H$ one recovers (12). Allowing variable metric can improve convergence rates, help to implicitly deal with certain constraints, or compensate the effect of ill-conditioning. Rather than simply giving a convergence result for a general choice of A_k, we handle, in Subsection 4.3, a detailed method to select these operators, using second-order information.

- Effectively solve structured problems as

$$\operatorname*{minimize}_{x_1\in H_1, x_2\in H_2} \; f(x_1, x_2) = g_1(x_1) + g_2(x_2) + h(x_1, x_2), \tag{15}$$

where g_1, g_2 are nonsmooth proper l.s.c functions and h is differentiable with Lipschitz gradient. One approach is the regularized Gauss-Seidel method, which exploits the fact that the variables are separated in the nonsmooth part of f, as considered in [17, 19, 20]. It consists in minimizing alternatively a regularized version of f with respect to each variable. In other words, it is an alternating proximal algorithm, of the form:

$$\left| \begin{array}{l} x_1^{k+1} \in \operatorname{prox}_{f(\cdot, x_2^k)}\left(x_1^k\right) \\ x_2^{k+1} \in \operatorname{prox}_{f(x_1^{k+1}, \cdot)}\left(x_2^k\right). \end{array} \right.$$

But this algorithm does not exploit the smooth nature of h. An alternative is to use an alternating minimization method which can deal with the nonsmooth character, while it benefits from the smooth features. An *Alternating Forward-Backward Method* considering variable metrics is presented below. A constant-metric version, namely the Proximal Alternating Linearized Minimization Algorithm, can be found in [21]. A forthcoming paper [22] deals with the same algorithm, called *Block Coordinate Variable Metric Forward-Backward*, with a non-cyclic way of selecting the variables to minimize. Nevertheless, our setting differs from the aforementioned works in the following ways:

- We allow more flexibility in the choice of parameters, accounting, in particular, for vanishing step sizes or unbounded eigenvalues for the metrics.
- We allow relative errors. Indeed, the computation of $\tilde{x}^k := x^k - A_k^{-1}\nabla h(x^k)$ and $x^{k+1} \in \operatorname{prox}_g^{A_k}\left(\tilde{x}^k\right)$ often require solving some subroutines, which may produce $\tilde{x}^k$ and x^{k+1} inexactly. To take these errors into account we introduce two sequences (r^k), (s^k) and consider

$$x^{k+1} - s^{k+1} \in \operatorname{prox}_g^{A_k}\left(x^k - A_k^{-1}\nabla h(x^k) + r^k\right). \tag{16}$$

Convergence of this method with errors is given in Theorem 6.

4.1. **The Alternating Forward-Backward (AFB) method.** Let $H_1, \ldots, H_p$ be Hilbert spaces, each H_i provided with its own inner product $\langle\cdot,\cdot\rangle_{H_i}$ and norm $\|\cdot\|_{H_i}$. If there is no ambiguity, we will just note $\|x_i\|$ instead of $\|x_i\|_{H_i}$. Set $H = \prod_{i=1}^{p} H_i$ and endow it with the inner product $\langle\cdot,\cdot\rangle = \sum_{i=1}^{p}\langle\cdot,\cdot\rangle_{H_i}$ and the associated norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$.

Consider the problem

$$\underset{x_i \in H_i}{\text{minimize}}\ f(x_1, \dots, x_p) = h(x_1, \dots, x_p) + \sum_{i=1}^{p} g_i(x_i), \tag{17}$$

where $h : H \to \mathbb{R}$ is continuously differentiable and each $g_i : H_i \to \mathbb{R} \cup \{+\infty\}$ is a lower-semicontinuous function. Moreover we suppose that there is $L \geq 0$ such that for each $(x_1, ..., x_p) \in H$ and $i \in \{1, ..., p\}$, the application

$$x \in H_i \mapsto h(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_p) \tag{18}$$

has a L-Lipschitz continuous gradient. We shall present an algorithm that generates sequences converging to critical points of f. The sequences will be updated cyclically, meaning that given $(x_1^k, ..., x_p^k)$, we start by updating the first variable x_1^k into x_1^{k+1}, and then we consider $(x_i^{k+1}, x_2^k, ..., x_p^k)$ to update the second variable, and so on. In order to have concise and clear notations, throughout this section we shall denote:

$$X^k := (x_1^k, ..., x_p^k) \text{ and } X_i^k := (x_1^{k+1}, ..., x_{i-1}^{k+1}, x_i^k, ..., x_p^k). \tag{19}$$

Observe that $X_1^k = X^k$ and that we can write $X_{p+1}^k = X^{k+1}$.

Let us now present the Alternate Forward-Backward (AFB) algorithm. As said before, it consists in doing a forward-backward step relatively to each variable, taking in account a possibly different metric. Then for all $i \in \{1, ..., p\}$, consider a sequence $(A_{i,k}) \subset \mathcal{S}_{++}(H_i)$ which will model the metrics. Given a starting point $X^0 \in H$, the AFB algorithm generates a sequence (X^k) by taking for all $k \in \mathbb{N}$ and $i \in \{1, ..., p\}$

$$\textbf{(AFB)} \quad x_i^{k+1} \in \text{prox}_{g_i}^{A_{i,k}} \left(x_i^k - A_{i,k}^{-1} \nabla_i h(X_i^k) \right). \tag{20}$$

We shall consider some hypotheses on the operators $A_{i,k}$. Define $\alpha_k = \min_{i=1..p} \alpha(A_{i,k})$ and $\beta_k := \max_{i=1..p} |||A_{i,k}|||$, which give bounds on the spectral values of $(A_{i,k})_{i=1..p}$. We make the following assumptions:

$$\textbf{(HP)} \quad \begin{array}{lll} 1.\ \text{There exists } \underline{\alpha} > 0 \text{ such that } \alpha_k \geq \underline{\alpha} > L & & \\ 2.\ \frac{1}{\beta_k} \notin \ell^1 & 3.\ \sup_{k \in \mathbb{N}} \frac{\beta_k}{\alpha_{k+1}} < +\infty. & \end{array}$$

Remark 1. Here $\mathbf{HP}_1$ is a bound on the spectral values by the Lipschitz constant of the gradient of h, in order to enforce the descent property of the sequence. For operators of the form $\frac{1}{\lambda_{i,k}} id_{H_i}$, we recover the classical bound $L\lambda_{i,k} \leq L\bar{\lambda} < 1$. In [22], the authors prove that, with an additional convexity assumption on the g_i's, and boundedness of the parameters, one can consider $L\lambda_{i,k} \leq L\bar{\lambda} < 2$. Item $\mathbf{HP}_2$ states that the spectral values may diverge, but not too fast. Finally, $\mathbf{HP}_3$ can be seen as an hypothesis on the variations of the extreme spectral values of the chosen operators. It clearly holds for instance if β_k is bounded. It is also sufficient to assume that the condition numbers

$$\kappa_i^k := \frac{|||A_i^k|||}{\alpha(A_i^k)}$$

are bounded, with also $\min\left\{ \frac{\alpha_k}{\alpha_{k+1}}, \frac{\beta_k}{\beta_{k+1}} \right\}$ remaining bounded.

Remark 2. Even if ∇h is globally Lipschitz continuous, L is not the Lipschitz constant of ∇h but a common Lipschitz constant for the functions defined in (18). As a consequence the partial gradients $\nabla_i h$ are $\sqrt{p}L$-Lipschitz continuous while ∇h is pL-Lipschitz. This allows us to have a better bound in $\mathbf{HP}_1$ which is of particular

importance in the applications (see Section 5). In [21], the authors give a more precise analysis: at each substep X_i^k of the algorithm, they consider $L_{i,k}$ as the Lipschitz constant of the gradient of $x \in H_i \mapsto h(x_1^{k+1}, ..., x_{i-1}^{k+1}, x, x_{i+1}^k, ..., x_p^k)$. Then they take step sizes equal to $\lambda_{i,k} = \frac{\epsilon_i}{L_{i,k}}$ where $\epsilon_i < 1$ is a fixed non-negative constant. This approach can be related to the one in [18, 22]. However, they suppose *a priori* that the values $L_{i,k}$ remain bounded. It would be interesting to know if it is possible to combine the two approaches (a variable Lipschitz constant and vanishing step sizes).

4.2. **The AFB method with errors.** In order to allow for approximate computation of the descent direction or the proximal mapping, we go further by considering an inexact AFB method. We introduce the sequences (r_i^k) and (s_i^k) for $i \in \{1, ..., p\}$ which correspond respectively to errors arising at the explicit and implicit steps relatively to the variable x_i. The *AFB method with Errors* is computed from an initial $(x_1^0, ..., x_p^0) \in H$ by

$$\textbf{(AFBE)} \qquad \begin{aligned} y_i^{k+1} &\in \operatorname{prox}_{g_i}^{A_{i,k}} \left(x_i^k - A_{i,k}^{-1} \nabla_i h(X_i^k) + r_i^k \right), \\ x_i^{k+1} &= y_i^{k+1} + s_i^{k+1}. \end{aligned}$$

We do specific hypothesis on the errors in view to guarantee the convergence of the method. Observe in particular that we do not assume a priori that the errors converge to zero:

(HE) There exists $\sigma \in [0, +\infty[, \rho \in]0, 1]$ with $\frac{\sigma+1}{\rho} < \underline{\alpha} L^{-1}$ such that
1. $\|S_i^k\| \leq \frac{\sigma}{2} \|y_i^{k+1} - y_i^k\|$, with S_i^k defined from (s_i^k) as in (19),
2. $\|r_i^k\| \leq \frac{\sigma}{2} \|y_i^{k+1} - y_i^k\| + \mu_k$, where $\mu_k \geq 0$ with $\mu_k \in \ell^1$,
3. $\langle r_i^k + s_i^k, y_i^{k+1} - y_i^k \rangle_{A_{i,k}} \leq \frac{1-\rho}{2} \|y_i^{k+1} - y_i^k\|_{A_{i,k}}^2$.

This AFB algorithm (with errors) is related to the abstract descent method studied in Section 3. This is stated in the next proposition, whose proof is left in Appendix A.2.

Proposition 1. *Any sequence $Y^k = (y_1^k, ..., y_p^k)$ generated by the AFB algorithm with errors satisfies $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$.*

Given this result, one could directly apply Theorem 1 to obtain convergence of the sequence (Y^k) to a critical point of f. But this result would suffer from some drawbacks. First, we are expecting that (X^k) converges to a critical point, not (Y^k). So we should make the assumption that the errors $S^k := X^k - Y^k$ tend to zero. Moreover we would suppose that (Y^k) is f-precompact, while we may only have an access to (X^k). To handle this, we make the link between the asymptotic behaviour of (Y^k) and (X^k):

Proposition 2. *For any sequence generated by the AFB method with errors:*

1. *If (Y^k) has finite length, then so does (X^k).*
2. *If $(f(Y^k))$ is bounded from below then for all $i \in \{1..p\}$, $\|s_i^k\|$ and $\|r_i^k\|$ lie in ℓ^2. In particular (Y^k) and (X^k) share the same limit points.*
3. *(Y^k) is precompact if and only if $(f(Y^k))$ is bounded from below and (X^k) is precompact.*

Proof. Item 1 comes directly from $\mathbf{HE}_1$. To prove item 2, we use Proposition 1: from $\mathbf{H}_1$ and $\mathbf{H}_3(i)$ we have that

$$\underline{a} \|Y^{k+1} - Y^k\|^2 \leq f(Y^k) - f(Y^{k+1}), \tag{21}$$

hence $(f(Y^k))$ is a decreasing sequence. Then we can sum inequality (21) to obtain that

$$\underline{a}\sum_{k\in\mathbb{N}}\|Y^{k+1}-Y^k\|^2 \leq f(Y^0)-\inf_{k\in\mathbb{N}} f(Y^k) < +\infty. \tag{22}$$

Since we have $\|r_i^k\| \leq \frac{\sigma}{2}\|y_i^{k+1}-y_i^k\|+\mu_k$ where $\mu_k \in \ell^1$ and $\|y_i^{k+1}-y_i^k\| \leq \|Y^{k+1}-Y^k\|$ which is in ℓ^2, we deduce that $\|r_i^k\| \in \ell^2$, and the same holds for $\|s_i^k\|$. So the errors converge to zero and (X^k) and (Y^k) have the same limit points. Item 3 follows from item 2 and the following: suppose that we have a subsequence (Y^{n_k}) converging to some $Y^\infty = (y_1^\infty, ..., y_p^\infty) \in \mathcal{H}$. Since f is lower semi-continuous and $(f(Y^k))$ is decreasing, we have that $\inf_{k\in\mathbb{N}} f(Y^k)$ is bounded from below by $f(Y^\infty)$. □

An other disadvantage to the direct application of Theorem 1 is that it asks the f-precompactness of (Y^k). In some cases, precompactness of a sequence can be deduced using compact embeddings between Hilbert spaces. Sequences remaining in a sublevel set of an inf-compact function f are also precompact. However, f-precompactness is harder to obtain without further continuity assumption on f. Actually, both limit and f-limit points coincide whenever the parameters are bounded:

Proposition 3. *If either $\beta_k \leq \bar{\beta}$ or f is continuous on its domain, then (Y^k) is f-precompact if and only if it is precompact.*

Proof. Suppose that we have Y^{k_n} converging to Y^∞, and show that $f(Y^{k_n})$ converges also to $f(Y^\infty)$. Note that $f(Y^k)$ being decreasing and f lower semicontinuous, we know that Y^∞ must lie in the domain of f. If f is continuous on its domain the conclusion is immediate. On the other hand suppose that $\beta_k \leq \bar{\beta}$. Since h is continuous, we only need to verify that $\lim_{n\to+\infty} g_i(y_i^{k_n}) = g_i(y_i^\infty)$ for each $i \in \{1..p\}$. The lower-semicontinuity of g_i already gives us $g_i(y_i^\infty) \leq \liminf_{n\to\infty} g_i(y_i^{k_n})$, so we just have to prove that $\limsup_{n\to\infty} g_i(y_i^{k_n}) \leq g_i(y_i^\infty)$, following the ideas of [17].
Let $n \in \mathbb{N}^*$ and $k = k_n - 1$, using the definition of the proximal operator, we have

$$\begin{aligned} &g_i(y_i^{k+1}) + \frac{1}{2}\|y_i^{k+1} - y_i^k + A_{i,k}^{-1}\nabla_i h(Y_i^k + S_i^k) - r_i^k - s_i^k\|_{A_{i,k}}^2 \\ &\leq g_i(y_i^\infty) + \frac{1}{2}\|y_i^\infty - y_i^k + A_{i,k}^{-1}\nabla_i h(Y_i^k + S_i^k) - r_i^k - s_i^k\|_{A_{i,k}}^2, \end{aligned}$$

and the latter implies (using Cauchy-Schwartz and $|||A_{i,k}||| \leq \bar{\beta}$):

$$g_i(y_i^{k+1}) \leq g_i(y_i^\infty) + \frac{\bar{\beta}}{2}\|y_i^\infty - y_i^k\|^2 + \|y_i^\infty - y_i^{k+1}\|\left[\|\nabla_i h(Y_i^k + S_i^k)\| + \bar{\beta}\|r_i^k + s_i^k\|\right]. \tag{23}$$

Now recall that $y_i^{k+1} = y_i^{k_n}$ tends to y_i^∞ while $r_i^k + s_i^k$ goes to zero (see Proposition 2). Observe also that $\nabla_i h(Y_i^k + S_i^k)$ is bounded since it converges to $\nabla_i h(Y^\infty)$. Moreover, $\|y_i^\infty - y_i^k\|$ goes also to zero since we have

$$\|y_i^\infty - y_i^k\| \leq \|y_i^\infty - y_i^{k_n}\| + \|y_i^{k+1} - y_i^k\|,$$

with $y_i^{k_n} \to y_i^\infty$ and $\|y_i^{k+1} - y_i^k\| \in \ell^2$ (see (22)). Passing to the upper limit in (23) leads finally to $\limsup_{n\to+\infty} g_i(y_i^{k_n}) \leq g_i(y_i^\infty)$. □

As a direct consequence of Propositions 1, 2, 3 together with Theorem 1, we finally get our convergence result for the AFB algorithm with errors. It extends the results of

[22] (when taking a cyclic permutation on the variables) in two directions: the functions g_i need not be continuous on their domain, or the step sizes can tend to 0.

Theorem 6. *Let f be a KŁ function. Let (Y^k) be a precompact sequence generated by the AFB algorithm with errors, with* **(HP)** *and* **(HE)** *satisfied. Suppose that either β_k remains bounded, or that f is continuous on its domain. Hence, the sequence (X^k) has finite length and converges toward a critical point of f.*

Remark 3. In the particular case where $S^k \equiv 0$, we know furthemore that the sequence (X^k) is convergent with respect to f. This is no longer true in general if f is not continuous and $S^k \neq 0$. As a simple counterexample, take $f : x \in \mathbb{R} \mapsto |x|_0 \in \mathbb{R}$ where $|x|_0 = 0$ if $x = 0$, $|x|_0 = 1$ else. By taking as parameters $A_k \equiv 2id$, $r^k \equiv 0$, $s^k = \frac{1}{k}$ and $x^0 = 0$, it is easy to see, after applying the AFB algorithm, that $f(y^k) \equiv 0$ but $f(x^k) \equiv 1$.

An analog of the capture result in Theorem 2 can also be deduced:

Theorem 7. *Suppose that the KŁ property holds in a global minimum X^* of f. Let (X^k) be a sequence generated by the AFB algorithm with errors, satisfying* **(HP)** *and* **(HE)** *with $\mu_k \equiv 0$. Hence, there exist $\gamma > 0$ and $\eta > 0$ such that if $X^0 \in \underline{\Gamma}_\eta(X^*, \gamma)$, then (X^k) has finite length and converges to a global minimum of f.*

To prove this theorem, it suffices to use $Y^0 = X^0$, and to see at the end of the proof of Proposition 1 that $\mu_k = 0$ iff $\epsilon_k = 0$, where ϵ_k is the parameter involved in $\mathbf{H}_3$. Then, apply Theorem 2 together with Propositions 1 and 2.

4.3. **Variable metric: towards generalized Newton methods.** We focus here on the problem of minimizing a $C^{1,1}$ function $h : \mathbb{R}^N \to \mathbb{R}$ over a closed nonempty set $C \subset \mathbb{R}^N$. The AFB algorithm reduces in this case to a projected-gradient method, and allow us to compute in the explicit step a descent direction governed by a chosen metric A_k. As an example, take $h(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ with $A \in \mathcal{S}_{++}(\mathbb{R}^N)$. In the unconstrained case, the Newton method (that is taking $A_k \equiv A$) is known to solve in one single step the problem. If we add a constraint C it is easy to see that the Newton-projected method

$$x^{k+1} \in \operatorname{proj}_C^{A_k} \left(x^k - A_k^{-1}\nabla h(x^k)\right) \tag{24}$$

gives the minimum of h over C in one single step. For a general function h, (24) reduces to the minimisation over C of a quadratic model of h, as stressed in (14). One can see on this example that computing the proximal operator relatively to the metric A_n used in the explicit step (and not the ambient metric !) is of crucial importance in this method.

The spirit here is to use second-order information from h in order to improve the convergence of the method. In the unconstrained case, a popular choice of metric is given by Newton-like methods, where the metric at step k is induced by (an approximation of) the Hessian $\nabla^2 h(x^k)$. Since it is often impossible to know in advance whether or not the Hessian is uniformly elliptic at each x^k, a positive definite approximation has to be chosen.
We detail here a natural way to chose this positive definite $A_k \sim \nabla^2 h(x^k)$ in closed loop, and show that this method remains in the setting of Theorem 6. Since it generalizes the *Levenberg-Marquardt* method used in the convex case (see [42]) we will refer to the *Generalized Levenberg-Marquardt* method for this way of designing A_k. One of the interesting aspect of the method is that such a matrix can be defined even if h is only $C^{1,1}$ and not C^2, since the differentiability of ∇h is not necessary in Theorem 6.

Another interesting aspect is that the splitting approach led us to solve constrained minimization problems with a Newton-projected approach.

We set $\mathcal{S}_+(\mathbb{R}^N)$ the closed convex cone of nonnegative matrices. Consider the generalized Hessian of h, by taking the generalized Jacobian of ∇h in sense of Clarke. Given $x \in \mathbb{R}^N$ it is

$$\partial^2 h(x) := co\{\lim_{n\to+\infty} \nabla^2 h(x_n), \text{ where } \nabla h \text{ is differentiable at } x_n \text{ and } x_n \to x\}.$$

This set contains symmetric matrices bearing second-order information on h. Hence, the Generalized Levenberg-Marquardt method to compute $A_k \in \mathcal{S}_{++}(\mathbb{R}^N)$ from a given $x^k \in \mathbb{R}^N$ is the following : for $\varepsilon > 0$,

$$\begin{array}{l} \text{Take } H_k \in \partial^2 h(x^k), \\ \text{Project } P_k = \text{proj}_{\mathcal{S}_+(\mathbb{R}^N)}(H_k), \\ \text{Regularize } A_k = P_k + \varepsilon I_N. \end{array}$$

A globalized version of the method can be considered by taking step sizes ensuring descent. Then the following convergence result holds:

Proposition 4. *Let $f(x) := h(x) + \delta_C(x)$ be a KŁ function, where $C \subset \mathbb{R}^N$ is closed nonempty and h is differentiable with a L-Lipschitz gradient. Let $x^0 \in H$ and suppose that (x^k) is a bounded sequence generated by*

$$x^{k+1} \in \text{proj}_C^{A_k}\left(x^k - \lambda_k A_k^{-1} \nabla h(x^k)\right),$$

where A_k is selected with the Generalized Levenberg-Marquardt process detailed above, and the stepsizes λ_k satisfy:

$$0 < \lambda_k \leq \bar{\lambda} < \frac{\varepsilon}{L}, \quad \lambda_k \notin \ell^1 \quad \textit{and} \quad \sup_{k\in\mathbb{N}} \frac{\lambda_{k+1}}{\lambda_k} < +\infty.$$

Then the sequence has finite length and is converging to a critical point of f.

Proof. Start by observing that $\text{proj}_C^{A_k} = \text{proj}_C^{\lambda_k^{-1} A_k}$, so the algorithm falls in the setting of the AFB algorithm. According with the previous notations, ∇h being L-Lipschitz continuous implies that the sequence $(|||H_k|||)$ is bounded by L, and so $(|||P_k|||)$ remains bounded by $2L$. To conclude through Theorem 6 we just need to check the hypotheses **(HP)** on the parameters $\frac{1}{\lambda_k} A_k$. We have here $\alpha_k = \alpha(\frac{1}{\lambda_k} A_k) \geq \varepsilon {\lambda_k}^{-1} \geq \varepsilon \bar{\lambda}^{-1} > L$ and $\beta_k = |||\frac{1}{\lambda_k} A^k||| \leq (2L + \epsilon)\lambda_k^{-1}$. Thus $\mathbf{HP}_1$ is satisfied, while items $\mathbf{HP}_2$ and $\mathbf{HP}_3$ follows directly from the hypotheses made on (λ_k). Since the indicator function δ_C is continuous on its domain, the hypotheses of Theorem 6 are satisfied. □

This extends, in a way, results from the convex setting to the nonconvex one, enforcing moreover the strong convergence (see [42, Theorem 7.1]).

A drawback of this method is that the Hessian increases the complexity of implementation since a matrix must be inverted in the explicit step. An alternative is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update scheme (see [30],[43]), using only first-order information to compute the inverse of the Hessian. On the other hand, the implicit step gains also in complexity since one must project onto a constraint relatively to a given metric, which is nontrivial even for simple constraints. For linear constraints, a particular second-order model of the Hessian can be taken in order to reduce the implicit step in a trivial orthogonal projection step (see [43, 44, 45]).

Newton-like methods are expected to have good convergence rates in exchange for a more expensive implementation. An interesting question is whether one can obtain convergence rates beyond the results in Subsection 3.2, by exploiting, not only the KŁ

nature of the function, but also the specific properties of the matrices selected by the Generalized Levenberg-Marquardt process.

5. Applications

The framework presented in this paper is suitable for the numerical resolution of a wide variety of structured problems. Consider for instance the problems arising in image processing and data compression, which are generally semi-algebraic by nature [47, 48, 49]. Indeed, they generally involve the semi-algebraic *counting norm* $\|x\|_0 := \sharp\{i \mid x_i \neq 0\}$, whose proximal operator (the hard shrinkage operator, see [17]) is easily implementable. Feasability problems with semi-algebraic (eventually nonconvex) constraints are also well suited for the AFB method (see [46, 17]). The search for equilibria of nonlinear partial differential equations has already been tackled using the KŁ inequality [15]. It should now be improved by using splitting methods more adapted to the structure of the problem. Let us end by discussing in some detail the *sparse and low-rank matrix decomposition*, for which the AFB method is particularly well adapted, in view of its structure.

Sparse and low-rank matrix decomposition. The problem of recovering the sparse and low-rank components of a matrix arises naturally in various areas such as model selection in statistics or system identification in engineering (see [50] and references therein). Denote by $\|X\|_0$ the number of nonzero components of $X \in \mathcal{M}_{m,n}(\mathbb{R})$. Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and bounds $r, s \in \mathbb{N}$, the low-rank sparse matrix decomposition problem consists in finding $X, Y \in \mathcal{M}_{m,n}(\mathbb{R})$ such that $A = X + Y$ with $\operatorname{rank}(x) \leq r$ and $\|Y\|_0 \leq s$. Endowing $\mathcal{M}_{m,n}(\mathbb{R})$ with the Frobenius norm, this reduces to

$$\underset{X,Y \in \mathcal{M}_{m,n}(\mathbb{R})}{\text{minimize}} \ \delta_{\{\operatorname{rank} \cdot \leq r\}}(X) + \delta_{\{\|\cdot\|_0 \leq s\}}(Y) + \frac{1}{2}\|A - X - Y\|_F^2.$$

An approach to solve this problem consists in doing a convex relaxation of the objective function (see [51, 52]). The sparsity and low-rank properties are obtained by minimizing the ℓ^1 and nuclear norms, respectively (see [53]).
The KŁ framework is well adapted to the original nonconvex (but semialgebraic!) problem and offers convergent numerical methods. Moreover, the AFB method is well suited for its structure in separated variables involving smooth and nonsmooth parts. It leads to an *Alternating Averaged Projected Method*: given (X_0, Y_0), take (λ_k), (μ_k) with $0 < \underline{\tau} \leq \lambda_k,\ \mu_k \leq \bar{\tau} < 1$. For $k \geq 0$, define

$$\begin{aligned} X^{k+1} &\in \operatorname{proj}_{\{\operatorname{rank} \cdot \leq r\}}(\lambda_k(A - Y^k) + (1 - \lambda_k)X^k), \\ Y^{k+1} &\in \operatorname{proj}_{\{\|\cdot\|_0 \leq s\}}(\mu_k(A - X^{k+1}) + (1 - \mu_k)Y^k). \end{aligned}$$

Projection onto $\{\operatorname{rank} \cdot \leq r\}$ can be done using the Singular Value Decomposition (see Eckart-Young's Theorem). To project onto $\{\|\cdot\|_0 \leq s\}$, one simply sets all the coefficients to zero, except for the s largest ones (in absolute value). Theorem 7 guarantees convergence to the solution for sufficiently close initialization. This example illustrate the discussion in Remark 2: here we have $L = 1$, while if one consider the Lipschitz constant of the gradient of $(X, Y) \mapsto \frac{1}{2}\|A - X - Y\|_F^2$, we would have had $L = 2$, that is a strictly smaller upper bound for the parameters.

6. Concluding Remarks

We have given a unified way to handle various recent descent algorithms, and derived general convergence rate results in the KŁ framework. These are applicable to

potential future numerical methods. Some improvements have been explored, and a novel projected Newton-like method has been proposed.

A challenging task is to extend the present convergence analysis to algorithms that do not satisfy the sufficient decrease condition $\mathbf{H}_1$. This will allow to consider acceleration schemes like the ones studied in [54, 55, 56], or primal-dual methods based on a Lagrangian approach. A recent preprint [57] seems to be an interesting first attempt in this direction.

From the applications point of view, the *counting norm* $\|\cdot\|_0$ evoked in Section 5 has a natural extension to an infinite-dimensional functional setting, namely the measure of the support of a function u defined on some $\Omega \subset \mathbb{R}^N$. An interesting but challenging issue is to apply our algorithm to this extension in order to solve the problem of sparse-optimal control of partial differential equations. From the implementation point of view, it suffices to apply the one-dimensional hard shrinkage operator at each point. Nevertheless, the verification of the KŁ inequality for this function has not been established and will probably rely on sophisticated arguments concerning the geometry of Hilbert spaces. Then, there is the natural question whether this approach is more efficient than those using the L^1 norm (see, for instance, [58]).

Finally, it is worth mentioning that the results in Section 3 remain true in the more general context of a normed space, adapting the definition of *subdifferential* and *lazy slope* in an obvious manner.

Acknowledgements : The authors thank H. Attouch for useful remarks. They would also like to thank the anonymous reviewer for his careful reading and constructive comments.

Appendix A. Appendix

A.1. **Proofs of Theorems 1 and 2.** The argument is a straightforward adaptation of the ideas in the proof of [17, Lemma 2.6]. One first proves:

Lemma 2. *Let* $\mathbf{H}_1$ *and* $\mathbf{H}_2$ *hold and fix* $k \in \mathbb{N}$. *If* x^k *and* x^{k+1} *belong to* $\underline{\Gamma}_\eta(x^*, \delta)$, *then*

$$2\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + \frac{1}{a_k b_k}\left[\varphi(f(x^k) - f(x^*)) - \varphi(f(x^{k+1}) - f(x^*))\right] + \epsilon_k. \tag{25}$$

For the next results, we introduce the following auxiliary property (automatically fulfilled under the hypotheses of Theorems 1 and 2), which includes a *stability* of the sequence $(x^k)_{k\in\mathbb{N}}$ with respect to the point x^*, along with a sufficiently close initialization.

$\underline{\mathbf{S}(x^*, \delta, \rho)}$: There exist $\delta > \rho > 0$ such that

i) For each $k \in \mathbb{N}$, if $x^0, \ldots, x^k \in \underline{\Gamma}_\eta(x^*, \rho)$, then $x^{k+1} \in \underline{\Gamma}_\eta(x^*, \delta)$;
ii) The initial point x^0 belongs to $\Gamma_\eta(x^*, \rho)$ and

$$\|x^* - x^0\| + 2\sqrt{\frac{f(x^0) - f(x^*)}{a_0}} + M\varphi(f(x^0) - f(x^*)) + \sum_{i=1}^{+\infty} \epsilon_i < \rho. \tag{26}$$

Then, we have the following estimation:

Lemma 3. *Let* $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$ *and* $\mathbf{S}(x^*,\delta,\rho)$ *hold, and note* $M=\sup_{k\in\mathbb{N}^*}\frac{1}{a_k b_k}<+\infty$. *Then, for all* $K\in\mathbb{N}^*$, *we have* $x^K\in\underline{\Gamma}_\eta(x^*,\rho)$ *and*

$$\sum_{k=1}^{K}\|x^{k+1}-x^k\|+\|x^{K+1}-x^K\|\leq\|x^1-x^0\|+M\big[\varphi(f(x^1)-f(x^*))-\varphi(f(x^{K+1})-f(x^*))\big]+\sum_{k=1}^{K}\epsilon_k.$$

The basic asymptotic properties are given by the following result:

Proposition 5. *Let* $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$ *and* $\mathbf{S}(x^*,\delta,\rho)$ *hold. Then* $x^k\in\underline{\Gamma}_\eta(x^*,\rho)$ *for all* k *and converges to some* $\overline{x}$ *lying in the closed ball* $\overline{B(x^*,\rho)}$. *Moreover* $\sum_{k=1}^{\infty}\|x^{k+1}-x^k\|<\infty$, $\liminf_{k\to\infty}\|\partial f(x^k)\|_-=0$, *and* $f(\overline{x})\leq\lim_{k\to\infty}f(x^k)=f(x^*)$.

Proof. Capture, convergence and finite length follow from Lemma 3 and $\mathbf{H}_3$. Next, since $(b_k)\notin\ell^1$ and $\sum_{k=1}^{\infty}b_{k+1}\|\partial f(x^k)\|_-\leq\sum_{k=1}^{\infty}\|x^{k+1}-x^k\|+\sum_{k=1}^{\infty}\varepsilon_{k+1}<\infty$, we obtain $\liminf_{k\to\infty}\|\partial f(x^k)\|_-=0$. Finally, observe that that $\lim_{k\to\infty}f(x^k)$ exists because $f(x^k)$ is decreasing and bounded from below by $f(x^*)$ and the lower-semicontinuity of f implies $f(\overline{x})\leq\lim_{k\to\infty}f(x^k)$. If $\lim_{k\to\infty}f(x^k)=\beta>f(x^*)$, the KŁ inequality and the fact that φ' is decreasing imply $\varphi'(\beta-f(x^*))\|\partial f(x^k)\|_-\geq\varphi'(f(x^k)-f(x^*))\|\partial f(x^k)\|_-\geq 1$ for all $k\in\mathbb{N}$, which is impossible because $\liminf_{k\to\infty}\|\partial f(x^k)\|_-=0$. Whence $\beta=f(x^*)$. □

We are now in position to complete the proofs of Theorems 1 and 2.

Proof of Theorem 1 Let $x^{n_k}\to x^*$ with $f(x^{n_k})\to f(x^*)$ as $k\to\infty$. Since $f(x^k)$ is nonincreasing and admits a limit point, we deduce that $f(x^k)\downarrow f(x^*)$. In particular, we have $f(x^*)\leq f(x^k)$ for all $k\in\mathbb{N}$. The function f satisfies the KŁ inequality on $\Gamma_\eta(x^*,\delta)$ with desingularizing function φ. Let $K_0\in\mathbb{N}$ be sufficiently large so that $f(x^K)-f(x^*)<\min\{\eta,\underline{a}\delta^2\}$, and pick $\rho>0$ such that $f(x^K)-f(x^*)<\underline{a}(\delta-\rho)^2$. Hence, $f(x^*)\leq f(x^{k+1})<f(x^*)+\eta$ for all $k\geq K$ and

$$\|x^{k+1}-x^k\|\leq\sqrt{\frac{f(x^k)-f(x^{k+1})}{a_k}}\leq\sqrt{\frac{f(x^K)-f(x^*)}{\underline{a}}}<\delta-\rho,$$

which implies part i) of $\mathbf{S}(x^*,\delta,\rho)$. Now take $K\geq K_0$ such that

$$\|x^*-x^K\|+2\sqrt{\frac{f(x^K)-f(x^*)}{a_{n_K}}}+M\varphi(f(x^K)-f(x^*))+\sum_{k=K+1}^{+\infty}\epsilon_k<\rho.$$

The sequence $(y^k)_{k\in\mathbb{N}}$ defined by $y^k=x^{K+k}$ for all $k\in\mathbb{N}$ satisfies the hypotheses of Proposition 5. Finally, since the whole sequence $(y^k)_{k\in\mathbb{N}}$ is f-convergent toward x^* and $\liminf_{k\to\infty}\|\partial f(y^k)\|_-=0$, we conclude that x^* must be critical using Lemma 1.

Proof of Theorem 2 Since f has the KŁ property in x^*, there is a strict local upper level set $\Gamma_\eta(x^*,\delta)$ where the KŁ inequality holds with φ as a desingularizing function. Take $\rho<\frac{3}{4}\delta$ and then $\gamma<\frac{1}{3}\rho$. If necessary, shrink η so that $2\sqrt{\frac{\eta}{\underline{a}}}+M\varphi(\eta)<\frac{2\rho}{3}$. This is possible since φ is continuous in 0 with $\varphi(0)=0$. Let $x^0\in\underline{\Gamma}_\eta(x^*,\gamma)\subset\underline{\Gamma}_\eta(x^*,\rho)$. It suffices to verify that $\mathbf{S}(x^*,\delta,\rho)$ is fulfilled and use Proposition 5. For i), let us suppose that $x^0,\ldots,x^k$ lie in $\underline{\Gamma}_\eta(x^*,\rho)$ and prove that $x^{k+1}\in\underline{\Gamma}_\eta(x^*,\delta)$. Since x^* is a global minimum, from $\mathbf{H}_1$ and the fact that $\big(f(x^k)\big)_{k\in\mathbb{N}}$ is decreasing, we have

$$f(x^*)+\underline{a}\|x^{k+1}-x^k\|^2\leq f(x^{k+1})+\underline{a}\|x^{k+1}-x^k\|^2\leq f(x^k)\leq f(x^0)<f(x^*)+\eta.$$

It follows that $\|x^{k+1}-x^*\| \leq \|x^{k+1}-x^k\| + \|x^k - x^*\| < \sqrt{\frac{\eta}{\underline{a}}} + \rho < \frac{4}{3}\rho < \delta$, and so $x^{k+1} \in \underline{\Gamma}_\eta(x^*, \delta)$. Finally, we have

$$\|x^0 - x^*\| + 2\sqrt{\frac{f(x^0)-f(x^*)}{a_0}} + M\varphi(f(x^0)-f(x^*)) < \frac{1}{3}\rho + 2\sqrt{\frac{\eta}{\underline{a}}} + M\varphi(\eta) < \rho,$$

which is precisely ii).

A.2. **Proof of Proposition 1.**

Proof. Since $X_i^k = Y_i^k + S_i^k$, we can rewrite the algorithm as

$$y_i^{k+1} \in \operatorname{prox}_{g_i}^{A_{i,k}}(y_i^k - A_{i,k}^{-1}\nabla_i h(Y_i^k + S_i^k) + r_i^k + s_i^k). \tag{27}$$

We start by showing that $\mathbf{H}_1$ is satisfied.

Let $i = 1..p$ be fixed. Using the definition of the proximal operator $\operatorname{prox}_{g_i}^{A_{i,k}}$ in (27) and developing the squared norms gives

$$\begin{aligned} &g_i(y_i^k) - g_i(y_i^{k+1}) \\ &\geq \frac{1}{2}\|y_i^{k+1} - y_i^k\|_{A_{i,k}}^2 + \langle y_i^{k+1} - y_i^k, \nabla_i h(Y_i^k + S_i^k)\rangle - \langle y_i^{k+1} - y_i^k, r_i^k + s_i^k\rangle_{A_{i,k}}. \end{aligned} \tag{28}$$

Using $\mathbf{HE}_3$ in (28), the latter results in

$$g_i(y_i^k) - g_i(y_i^{k+1}) \geq \frac{1}{2}\|y_i^{k+1} - y_i^k\|_{\rho A_{i,k}}^2 + \langle y_i^{k+1} - y_i^k, \nabla_i h(Y_i^k + S_i^k)\rangle. \tag{29}$$

For fixed $k \in \mathbb{N}$ and $i = 1, \ldots, p$, introduce the function

$$\tilde{h}_{i,k} : y_i \in H_i \mapsto (y_1^{k+1}, .., y_{i-1}^{k+1}, y_i, y_{i+1}^k, .., y_p^k) \in \mathbb{R} \tag{30}$$

which satisfies $\tilde{h}_{i,k}(y_i^k) = h(Y_i^k)$, $\tilde{h}_{i,k}(y_i^{k+1}) = h(Y_{i+1}^k)$ and $\nabla\tilde{h}_{i,k}(y_i^k) = \nabla_i h(Y_i^k)$. Applying the descent lemma to $\tilde{h}_{i,k}$, we obtain

$$h(Y_{i+1}^k) - h(Y_i^k) - \langle y_i^{k+1} - y_i^k, \nabla_i h(Y_i^k)\rangle \leq \frac{L}{2}\|y_i^{k+1} - y_i^k\|^2. \tag{31}$$

Then, combining (29) and (31) we get

$$\begin{aligned} &g_i(y_i^k) - g_i(y_i^{k+1}) + h(Y_i^k) - h(Y_{i+1}^k) \\ &\geq \frac{1}{2}\|y_i^{k+1} - y_i^k\|_{\rho A_{i,k} - L id_{H_i}}^2 + \langle y_i^{k+1} - y_i^k, \nabla_i h(Y_i^k + S_i^k) - \nabla_i h(Y_i^k)\rangle, \end{aligned} \tag{32}$$

where $\rho A_{i,k} - L id_{H_i}$ remains coercive, since $\rho\alpha_k > L$. Using successively the Cauchy-Schwartz inequality, the Lipschitz property of $\nabla_i h$ (see Remark 2) and $\mathbf{HE}_1$, one gets

$$\begin{aligned} \langle y_i^{k+1} - y_i^k, \nabla_i h(Y_i^k + S_i^k) - \nabla_i h(Y_i^k)\rangle &\geq -\|y_i^{k+1} - y_i^k\|\|\nabla_i h(Y_i^k + S_i^k) - \nabla_i h(Y_i^k)\| \\ &\geq -L\ \|y_i^{k+1} - y_i^k\|\|S_i^k\| \geq -\frac{\sigma L}{2\sqrt{p}}\|y_i^{k+1} - y_i^k\|^2. \end{aligned}$$

Inserting this estimation in (32) we deduce that

$$g_i(y_i^k) - g_i(y_i^{k+1}) + h(Y_i^k) - h(Y_{i+1}^k) \geq \frac{1}{2}\|y_i^{k+1} - y_i^k\|_{\rho A_{i,k} - L(\sigma\frac{\sqrt{p}}{p}+1) id_{H_i}}^2. \tag{33}$$

We can now conclude by summing all these inequalities for $i = 1, \dots, p$:

$$f(Y^k) - f(Y^{k+1}) = \sum_{i=1}^{p} g_i(y_i^k) - g_i(y_i^{k+1}) + h(Y_i^k) - h(Y_{i+1}^k) \tag{34}$$
$$\geq \frac{1}{2} \sum_{i=1}^{p} \|y_i^{k+1} - y_i^k\|^2_{\rho A_{i,k} - L(\sigma \frac{\sqrt{p}}{p} + 1) id_{H_i}}$$

so $\mathbf{H}_1$ is fulfilled with $a_k = \frac{\rho \alpha_k - L(\sigma \frac{\sqrt{p}}{p} + 1)}{2}$.

To prove $\mathbf{H}_2$, fix $i = 1, \dots, p$ and use Fermat's first order condition in (27) to get:

$$0 \in \partial g_i(y_i^{k+1}) + \left\{ A_{i,k}(y_i^{k+1} - y_i^k) - A_{i,k}(r_i^k + s_i^k) + \nabla_i h(Y_i^k + S_i^k) \right\} \tag{35}$$

Define $w_i^{k+1} := \nabla_i h(Y^k) - \nabla_i h(Y_i^k + S_i^k) - A_{i,k}(y_i^{k+1} - y_i^k) + A_{i,k}(r_i^k + s_i^k)$ which lies in $\partial g_i(y_i^{k+1}) + \nabla_i h(Y^{k+1})$, by (35). The triangle inequality gives

$$\|w_i^{k+1}\| \leq \beta_k \left(\|y_i^{k+1} - y_i^k\| + \|r_i^k\| + \|s_i^k\| \right) + \|\nabla_i h(Y_i^k + S_i^k) - \nabla_i h(Y^{k+1})\|, \tag{36}$$

where we use the error estimations from **(HE)**

$$\|r_i^k\| + \|s_i^k\| \leq \sigma \|y_i^{k+1} - y_i^k\| + \mu_k, \tag{37}$$

and the $\sqrt{p}L$-Lipschitz continuity of $\nabla_i h$:

$$\|\nabla_i h(Y_i^k + S_i^k) - \nabla_i h(Y_i^k)\| \leq \sqrt{p} L \|Y_i^k - Y^{k+1} + S_i^k\| \tag{38}$$
$$\leq \sqrt{p} L \|Y^{k+1} - Y^k\| + \sqrt{p} L \sigma \|y_i^{k+1} - y_i^k\|.$$

Combining (36), (37) and (38) leads to

$$\|w_i^{k+1}\| \leq (\beta_k(1+\sigma) + \sqrt{p} L \sigma) \|y_i^{k+1} - y_i^k\| + \sqrt{p} L \|Y^{k+1} - Y^k\| + \beta_k \mu_k. \tag{39}$$

Define now $W^{k+1} := (w_1^{k+1}, ..., w_p^{k+1}) \in \partial f(Y^{k+1})$ (recall the definition of w_i^{k+1}). Then through the sum over $i = 1..p$ of inequality (39) we have (using $\sqrt{p} \leq p \leq p^2$)

$$\|W^{k+1}\| \leq \sum_{i=1}^{p} \|w_i^{k+1}\| \leq p \beta_k \mu_k + p^2 (\beta_k + L)(1+\sigma) \|Y^{k+1} - Y^k\|.$$

Hence $\mathbf{H}_2$ is verified with $b_{k+1} = \frac{1}{p^2(1+\sigma)(\beta_k + L)}$ and $\epsilon_{k+1} = \frac{\beta_k \mu_k}{p(1+\sigma)(\beta_k + L)}$.

Now we just need to check that the hypotheses $\mathbf{H}_3$ are satisfied with our hypotheses on α_k, β_k and μ_k. Clearly $\mathbf{H}_3(i)$ holds since we've supposed that $\alpha_k \geq \underline{\alpha} > (\sigma \frac{\sqrt{p}}{p} + 1) \frac{L}{\rho}$. Then $\mathbf{H}_3(ii)$ asks that $b_k \notin \ell^1$, which is equivalent to $\frac{1}{\beta_k + L} \notin \ell^1$ in our context. This holds since we've supposed that $\frac{1}{\beta_k} \notin \ell^1$. Hypothese $\mathbf{H}_3(iii)$ is satisfied because $\frac{\beta_k}{\alpha_{k+1}}$ is supposed to be bounded. Finally, $\mathbf{H}_3(iv)$ asks the summability of $\frac{\beta_k \mu_k}{\beta_k + L}$ which is bounded by $\mu_k \in \ell^1$. □

References

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

[2] Kurdyka K.: On gradients of functions definable in o-minimal structures, Ann. Inst. Fourier 48, 769–783 (1998)

[3] Łojasiewicz, S.: Sur la géométrie semi- et sous-analytique. Ann. Inst. Fourier 43, 1575–1595 (1993)

[4] Kurdyka, K., Parusiński, A.: $\mathbf{w}_f$-stratification of subanalytic functions and the Łojasiewicz inequality. C. R. Acad. Paris 318, 129–133 (1994)

[5] Bolte J., Daniilidis A., Lewis A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17, 1205–1223 (2006)
[6] Simon L.: Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems, Ann. Math. 118, 525–571 (1983)
[7] Huang, S.-Z., Takáč, P.: Convergence in gradient-like systems which are asymptotically autonomous and analytic. Nonlinear Anal., Ser. A Theory Methods 46, 675–698 (2001)
[8] Chill R., Jendoubi M.A.: Convergence to steady states in asymptotically autonomous semilinear evolution equations, Nonlinear Anal. 53, 1017–1039 (2003)
[9] Haraux A., Jendoubi M.A.: Convergence of solutions of second-order gradient-like systems with analytic nonlinearities. J. Differential Equations 144, no. 2, 313–320 (1998)
[10] Haraux A.: A hyperbolic variant of Simon's convergence theorem. Evolution equations and their applications in physical and life sciences (Bad Herrenalb, 1998), Lecture Notes in Pure and Appl. Math., vol. 215, pp. 255-264. Dekker, New York (2001)
[11] Baudoin, L., Salomon, J.: Constructive solution of a bilinear optimal control problem for a Schrödinger equation. Syst Contr Lett. 57, no. 6, 453–464 (2008).
[12] Absil, P.-A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. 16, 531–547 (2005)
[13] Attouch H., Bolte J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Math. Program. Ser. B 116, 5–16 (2009)
[14] Bolte J., Daniilidis A., Ley O., Mazet L.: Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity, Trans. Amer. Math. Soc. 362, 3319–3363 (2010)
[15] B. Merlet, M. Pierre: Convergence to equilibrium for the backward Euler scheme and applications, Commun. Pure Appl. Anal 9, 685–702, (2010).
[16] Noll D.: Convergence of Non-smooth Descent Methods Using the Kurdyka-Lojasiewicz Inequality. J Optim Theory Appl., 160, 553–572 (2014).
[17] Attouch H., Bolte J., Svaiter B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program. 137, no. 1-2, 91–129 (2013)
[18] Chouzenoux E., Pesquet J.C., Repetti A.: Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. To appear in J. Optim. Theory Appl. (2013).
[19] Attouch H., Bolte J., Redont P., Soubeyran A.: Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Łojasiewicz inequality, Math. Oper. Res. 35(2), 438–457 (2010)
[20] Xu Y., Yin W.: A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion, SIAM J. Imaging Sciences, Vol. 6, No. 3, pp. 1758-1789, (2013).
[21] Bolte J., Sabach S., Teboulle M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Programm. (2013)
[22] Chouzenoux E., Pesquet J.C., Repetti A.: A Block Coordinate Variable Metric Forward-Backward Algorithm, submitted (2014). Preprint available online at `www.optimization-online.org/DB_HTML/2013/12/4178.html`
[23] Li D., Pang L.-P., Chen S.: A proximal alternating linearization method for nonconvex optimization problems. Optimization Methods & Software, 29, no. 4, (2014).
[24] Bolte J., Daniilidis A., Lewis A., Shiota M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. 18, no. 2, 556–572 (2007)
[25] van den Dries L.: Tame topology and o-minimal structures, Bull. of the AMS 37(3), 351–357 (2000)
[26] van den Dries L., Miller C.: Geometric categories and o-minimal structures, Duke Math. J. 84, 497–540 (1996).
[27] Haraux A., Jendoubi M.A.: The Łojasiewicz gradient inequality in the infinite dimensional Hilbert space framework, J. of Func. Anal. 260, No. 9, 2826–2842 (2010)
[28] Chill R.: The Łojasiewicz-Simon gradient inequality in Hilbert spaces, Proceedings of the 5th European-Maghrebian Workshop on Semigroup Theory, Evolution Equations, and Applications (M. A. Jendoubi, ed.), 25–36 (2006)
[29] H. Attouch, G. Buttazo, G. Michaille: Variational Analysis in Sobolev and BV Spaces. MPS-SIAM Series on Optimization, Springer, (2008).
[30] Bertsekas D.: Nonlinear programming, Athena Scientific, Belmont MA, 1999.

[31] Rockafellar R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optimization 14, no. 5, 877–898 (1976).
[32] Ferris M.: Finite termination of the proximal point algorithm, Math. Program. 50, 359-366 (1991)
[33] Peypouquet J.: Asymptotic Convergence to the Optimal Value of Diagonal Proximal Iterations in Convex Minimization, J. Convex Anal. 16, no. 1, 277–286 (2009)
[34] Chill R., Fiorenza A.: Convergence and decay rate to equilibrium of bounded solutions of quasilinear parabolic equations, J. Differential Equations, 228, 611–632 (2006)
[35] Cauchy, A.-L., Méthode générale pour la résolution des systèmes d'équations simultanées. C. R. Acad. Sci. Paris, 25, 536–538 (1847).
[36] B. Martinet, Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle 4, Sér. R-3, 154–158 (1970).
[37] H. Brézis, P.-L. Lions: Produits infinis de résolvantes. Israel J. Math. 29, no. 4, 329–345 (1978).
[38] P.L. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis, 16, 964–979, (1979).
[39] G. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. 72, no. 2, 383–390 (1979).
[40] F. Alvarez, J. Bolte, O. Brahic, Hessian Riemannian gradient flows in convex programming. SIAM J. Control Optim. 43, no. 2, 477–501 (2004).
[41] F. Alvarez, J. López, H. Ramírez, Interior proximal algorithm with variable metric for second-order cone programming: applications to structural optimization and support vector machines. Optim. Methods Softw. 25, no. 4-6, 859–881 (2010).
[42] H. Attouch, B.F. Svaiter: A continuous dynamical Newton-like approach to solving monotone inclusions. SIAM J. Optim., Vol. 49, No. 2, pp. 574-598, (2011).
[43] S. Sra, S. Nowozin, S. J. Wright: Optimization for machine learning. MIT Press, (2011).
[44] E. M. Gafni, D. P. Bertsekas: Two-metric projection methods for constrained optimization. SIAM J. on Control and Optim., Vol. 22, No. 6, pp. 936-964, (1984).
[45] D. P. Bertsekas: Projected Newton methods for optimization problems with simple constraints. SIAM J. on Control and Optim., Vol. 20, No. 2, pp. 221-246, (1982).
[46] A.S. Lewis, D.R. Luke, J. Malick: Local linear convergence for alternating and averaged non convex projections. Found. of Comp. Mat., Vol. 9, No. 4, pp. 485-513, (2009).
[47] D.L. Donoho: For most Large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. Comm. on Pure and Appl. Math., Vol. 59, Vol. 6, pp. 797-829, (2006).
[48] D.L. Donoho, J. Tanner: Counting the faces of randomly-projected hypercubes and orthants, with applications. Discrete Comput. Geom., Vol. 43, pp. 522-541, (2010).
[49] I. Daubechies, M. Defrise and C. De Mol: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math, Vol. 57, No. 11, pp. 1413-1457, (2004).
[50] Chandrasekaran V., Sanghavi S., Parillo P.A., Willsky A.S.: Rank-sparsity incoherence for matrix decomposition, SIAM J. Opt., Vol. 21, No. 2, pp. 572-596, (2011).
[51] Ganesh A., Lin Z., Wright J., Wu L., Chen M., Ma Y.: Fast algorithms for recovering a corrupted low-rank matrix., in Proceedings of the 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 213-216 (2009).
[52] Yuan X., Yang J.: Sparse and low-rank matrix decomposition via alternating direction method, Pacific Journal of Optimization, 9(1), 167-180, (2013).
[53] Recht B., Fazel B., Parillo P.A.: Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization, SIAM Rev., Vol. 52, pp. 471-501, (2007).
[54] Nesterov Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269, pp. 543–547 (1983)
[55] Beck A., Teboulle M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, SIAM J. on Imaging Sci., 2, pp. 183–202 (2009)
[56] A. Beck and M. Teboulle, Gradient-based algorithms with applications in signal recovery problems, In Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar Eds., pp. 33–88. Cambridge University Press (2010)
[57] Ochs P., Chen Y., Brox T., Pock T.: iPiano: Inertial Proximal Algorithm for Non-convex Optimization, submitted (2014). Preprint available online at `http://lmb.informatik.uni-freiburg.de/Publications/2014/OB14/`
[58] Casas E., Herzog R., Wachsmuth G.: Approximation of sparse controls in semilinear equations by piecewise linear functions, Numer. Math. 122, no. 4, 645-669 (2012)