SIAM J. OPTIM.
Vol. 29, No. 1, pp. 541–570

UNIFYING ABSTRACT INEXACT CONVERGENCE THEOREMS AND BLOCK COORDINATE VARIABLE METRIC IPIANO*

PETER OCHS†

Abstract. An abstract convergence theorem for a class of generalized descent methods that explicitly models relative errors is proved. The convergence theorem generalizes and unifies several recent abstract convergence theorems. It is applicable to possibly nonsmooth and nonconvex lower semicontinuous functions that satisfy the Kurdyka–Łojasiewicz (KL) inequality, which comprises a huge class of problems. Many of the recent algorithms that explicitly prove convergence using the KL inequality can be cast in the abstract framework of this paper and, therefore, the generated sequence converges to a stationary point of the objective function. Additional flexibility compared to related approaches is gained by a descent property that is formulated with respect to a function that is allowed to change along the iterations, a generic distance measure, and an explicit/implicit relative error condition with respect to finite linear combinations of distance terms. As an application of the gained flexibility, the convergence of a block coordinate variable metric version of iPiano (an inertial forward-backward splitting algorithm) is proved, which performs favorably on an inpainting problem with a Mumford–Shah-like regularization from image processing.

Key words. abstract convergence theorem, Kurdyka–Łojasiewicz inequality, descent method, relative errors, block coordinate method, variable metric method, inertial method, iPiano, inpainting, Mumford–Shah regularizer

AMS subject classifications. 32B20, 47J06, 49J52, 65K10, 90C06, 90C30, 65K05

DOI. 10.1137/17M1124085

1. Introduction.

The Kurdyka–Łojasiewicz (KL) inequality is key for the convergence analysis for nonsmooth and nonconvex optimization problems. Łojasiewicz introduced an early version of this inequality for analytic functions [37], which was extended to more general classes of smooth functions in [29, 38, 30] and to nonsmooth functions (that are definable in an o-minimal structure [20]) in [8, 9]. While it was originally used to study the asymptotic behavior of gradient-like systems [8, 24, 26, 31] and PDEs [17, 49], the KL inequality is also used for numerical methods such as the gradient method [1], proximal methods [3], projection or alternating minimization methods [4, 7]. A unifying and concise formulation of the key ingredients, which, combined with the KL inequality, lead to asymptotic convergence to a critical point and a trajectory with finite length (the accumulated distance between consecutive points of the sequence is finite), is proposed by Attouch, Bolte, and Svaiter [5], and further refined by Bolte, Sabach, and Teboulle [12] using a uniformization result for the KL inequality. These early developments revolutionized the study of numerical methods for nonsmooth nonconvex optimization problems.

In this work, we continue the abstract unification of the convergence analysis of algorithms for nonsmooth nonconvex optimization [5, 12]. Their convergence analysis is driven by two central assumptions: a *sufficient decrease condition* and a *relative error condition*. While they use the sufficient decrease condition on the objective function, [44] formulates conditions that apply to a global surrogate function of Lyapunov type, which allows the objective values also to increase locally. Note that

*Received by the editors April 3, 2017; accepted for publication (in revised form) September 14, 2018; published electronically February 19, 2019.
http://www.siam.org/journals/siopt/29-1/M112408.html
†Mathematical Optimization Group, Saarland University, 66123 Saarbruecken, Germany (ochs@math.uni-sb.de).

this idea is different from the majorization minimization principle [27], where in each iteration a majorizer of the objective is constructed and minimized, which usually leads to a descent of the actual objective values. In the KL context, this algorithmic strategy was used in [11, 45], and led to another abstract convergence result in [11] similar to that of [5]. The abstract conditions formulated in our paper contain those of [5, 12, 11, 44, 45] as special instances.

The relative error condition is justified by the fact that most algorithms require solving subproblems for which possibly inexact approaches are required. The condition reflects *relative inexact optimality conditions* [5], and is related to [28, 51, 50, 52]. In [11] the relative error condition is of explicit nature (see also [1, 42]), whereas in [5, 12, 44] it is implicit. The abstract convergence theorem in our paper comprises the explicit and the implicit formulation.

The sufficient decrease condition and the relative error condition depend rather on the structure of the algorithm than on fine properties of the objective function. Therefore, the parameters appearing in these conditions are tightly linked to properties of the algorithm such as the step size. While the abstract convergence conditions discussed so far rely on a constant choice of these parameters, Frankel, Garrigos, and Peypouquet [21] introduced a significantly more flexible parameter setting into these conditions. As a result, an alternating version of the variable metric forward-backward splitting algorithm is formulated and its convergence is proved, which opens the door for a nonsmooth and nonconvex version of the Levenberg–Marquardt algorithm. The conditions in our paper are formulated such that [21] appears as a special case.

Beyond the flexibility introduced in [21], in this paper, (1) we allow for a parametric function for which the sufficient decrease condition is required. This allows the objective or any surrogate relative for which decrease is measured to change along the iterations. We believe that this additional flexibility has significant potential, which in this paper is only rudimentarily explored in the context of an inertial variable metric method. (2) The relative error condition can be formulated with respect to a linear combination of finitely many distance terms, which seems to be essential for multistep methods [44, 43, 15, 36]. Finally, (3) all distances and the decrease in (1) are formulated using abstract distances. Of course, unless there is a closer relation between the abstract distance measure and the Euclidean metric, we have to content ourselves with a weaker convergence result. Nevertheless, we consider this as an essential step in generalizing the convergence results further, possibly to algorithms that use Bregman distances [16] without smoothness or strong convexity assumption. In the present paper, we use the abstract distance measures to restrict the Euclidean distance to blocks of coordinates, which leads (almost for free) to a block coordinate version of the inertial variable metric method iPiano. Without the variable metric aspect, the block coordinate inertial method was already proposed in [46], though as a result of a more explicit analysis.

So far, we have focused on abstract convergence results for nonsmooth nonconvex optimization problems. As mentioned above, there are many concrete algorithms that are proved to converge in such a general setting using the abstract conditions or an explicit verification of the convergence following the lines of the abstract convergence proof.

Convergence of the *gradient method* is proved in [1, 5], and has been extended to *proximal gradient descent* (also known as the forward-backward splitting method) [5], which applies to a class of problems that is given as the sum of a (possibly nonsmooth and nonconvex) function and a smooth (possibly nonconvex) function. Accelerations by means of a *variable metric* are considered in [18, 21], and in combination with a line-search procedure in [13]. The convergence of *proximal methods* is inspected in [3, 5, 10, 40], and an *alternating proximal method* is considered in [4]. Extensions to *block coordinate* methods are given, e.g., in [5] under the name regularized Gauss–Seidel method, which is actually a variable metric version of the block coordinate methods in [4, 6, 23]. The combination of the ideas of alternating proximal minimization and forward-backward splitting can be found in [12], where the algorithm is called *proximal alternating linearized minimization (PALM)*. For an extension that allows the metric to change in each iteration with a flexible order of the block iterations, we refer the reader to [19]. Convergence of a nonsmooth subgradient method is studied in [42, 25].

Another way of accelerating descent methods (instead of using a variable metric) is via so-called *inertial methods*. In convex optimization, some inertial or overrelaxation methods are known to be optimal [41]. Although it is hard to obtain sharp lower complexity bounds in the nonconvex setting, and hence to argue about optimal methods, experiments show favorable performance of inertial algorithms. In [44] an extension of inertial gradient descent (also known as the *heavy-ball method* or gradient descent with momentum), which includes an additional nonsmooth convex term in the objective function similar to forward-backward splitting, is analyzed in the KL framework. The proposed algorithm is called *iPiano* and shows good performance in applications. An earlier subsequential convergence proof of Polyak's heavy-ball method [47] without the KL inequality for smooth nonconvex functions is proposed in [54]. In [43, 15] the original problem class "nonsmooth convex plus smooth nonconvex" in [44] was extended to "nonsmooth nonconvex plus smooth nonconvex". In [15] (smooth and strongly convex) Bregman proximity functions are used in the update step. See [14] for a variant of this algorithm. A block coordinate version of iPiano or an *inertial variant of the proximal alternating linearized minimization* method was recently proposed as iPALM in [46]. A variable metric version of iPiano and iPALM—*block coordinate variable metric iPiano*—is proposed in this paper. The accelerated method in [35] is based on an *extrapolation* of the gradient similar to Nesterov's proximal gradient method instead of an inertial term. Liang, Fadili, and Peyré [36] pursue a unifying approach of the preceding methods by way of a *generic multistep method*. All of these inertial methods share the property that the sufficient decrease condition holds for a Lyapunov function instead of the actual objective function.

This concept is important beyond inertial methods. It is used to prove convergence of splitting methods for *composite problems* [33], *Douglas–Rachford splitting* [32], and *Peaceman–Rachford splitting* [34] for nonconvex optimization problems.[1]

Section 2 introduces the basic notation and results from (nonsmooth) variational analysis [48] and the Kurdyka–Łojasiewicz inequality. Section 3 formulates the basic conditions for the abstract convergence theorem, which is motivated by the results in [5, 21, 44, 12]. The gained flexibility of the conditions is compared to related work in section 3.1, and further discussed in section 3.2, where some future perspectives are also provided. Examples for the necessity of the generalizations are given in Appendix A. The convergence under the abstract conditions is proved in section 3.3. The flexibility that is gained is used in section 4 to prove convergence of a variable metric version of iPiano [44, 43] and in section 5 of a block coordinate variable metric version of iPiano. Several block coordinate, variable metric, and inertial versions of forward-backward splitting/iPiano are applied to an image inpainting problem in section 6, which emphasizes the importance of a variable metric and block coordinate methods.

[1] Unlike in the convex setting, the Douglas–Rachford splitting mentioned here requires one function in the objective to be smooth.

2. Preliminaries.

2.1. Notation and definitions. Throughout this paper, we will always work in a finite-dimensional Euclidean vector space $\mathbb{R}^N$ of dimension $N \in \mathbb{N}$, where $\mathbb{N} := \{1, 2, \ldots\}$. Define $\mathbb{Z} := \{\ldots, -1, 0, 1, \ldots\}$. The vector space is equipped with the standard Euclidean norm $\|\cdot\| := \|\cdot\|_2$ that is induced by the standard Euclidean inner product $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. If specified explicitly, we work in a metric induced by a symmetric positive definite matrix $A \in \mathbb{S}_{++}(N) \subset \mathbb{R}^{N\times N}$, represented by the inner product $\langle x, y\rangle_A := \langle Ax, y\rangle$ and the norm $\|x\|_A := \sqrt{\langle x, x\rangle_A}$. For $A \in \mathbb{S}_{++}(N)$ we define $\varsigma(A) \in \mathbb{R}$ as the largest value that satisfies $\|x\|_A^2 \geq \varsigma(A)\|x\|_2^2$ for all $x \in \mathbb{R}^N$.

As usual, we consider extended real-valued functions $f : \mathbb{R}^N \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$, that are defined on the whole space, with *domain* $\operatorname{dom} f := \{x \in \mathbb{R}^N \,|\, f(x) < +\infty\}$. A function is called *proper* if $\operatorname{dom} f \neq \emptyset$. We define the *epigraph* of the function f as $\operatorname{epi} f := \{(x, \mu) \in \mathbb{R}^{N+1} \,|\, \mu \geq f(x)\}$. We will also need to consider set-valued mappings $F : \mathbb{R}^N \rightrightarrows \mathbb{R}^M$ defined by the *graph*

$$\operatorname{Graph} F := \{(x, y) \in \mathbb{R}^N \times \mathbb{R}^M \,|\, y \in F(x)\},$$

where the domain of a set-valued mapping is given by $\operatorname{dom} F := \{x \in \mathbb{R}^N \,|\, F(x) \neq \emptyset\}$. For a proper function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ we define the set of *(global) minimizers* as

$$\arg\min f := \arg\min_{x\in\mathbb{R}^N} f := \{x \in \mathbb{R}^N \,|\, f(x) = \inf f\}, \qquad \inf f := \inf_{x\in\mathbb{R}^N} f(x).$$

The *Fréchet subdifferential* of f at $\bar{x} \in \operatorname{dom} f$ is the set $\widehat{\partial} f(\bar{x})$ of those elements $v \in \mathbb{R}^N$ such that

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

For $\bar{x} \notin \operatorname{dom} f$, we set $\widehat{\partial} f(\bar{x}) = \emptyset$. For convenience, we introduce *f-attentive convergence*: a sequence $(x^n)_{n\in\mathbb{N}}$ is said to *f-converge* to $\bar{x}$ if

$$x^n \to \bar{x} \quad \text{and} \quad f(x^n) \to f(\bar{x}) \quad \text{as } n \to \infty,$$

and we write $x^n \xrightarrow{f} \bar{x}$. The so-called *(limiting) subdifferential* of f at $\bar{x} \in \operatorname{dom} f$ is defined by

$$\partial f(\bar{x}) := \left\{ v \in \mathbb{R}^N \,\middle|\, \exists\, x^n \xrightarrow{f} \bar{x},\ v^n \in \widehat{\partial} f(x^n),\ v^n \to v \right\},$$

and $\partial f(\bar{x}) = \emptyset$ for $\bar{x} \notin \operatorname{dom} f$. A point $\bar{x} \in \operatorname{dom} f$ for which $0 \in \partial f(\bar{x})$ is a called a *critical point* or *stationary point*. As a direct consequence of the definition of the limiting subdifferential, we have the following closedness property:

$$x^n \xrightarrow{f} \bar{x},\ v^n \to \bar{v},\ \text{and for all } n \in \mathbb{N}\colon v^n \in \partial f(x^n) \quad \Longrightarrow \quad \bar{v} \in \partial f(\bar{x}).$$

Exercise 8.8 of [48] shows that at a point $\bar{x} \in \mathbb{R}^N$, for the sum of an extended-valued function g that is finite at $\bar{x}$ and a continuously differentiable (smooth) function f around $\bar{x}$, it holds that $\partial(g + f)(\bar{x}) = \partial g(\bar{x}) + \nabla f(\bar{x})$. Moreover for a function $f : \mathbb{R}^N \times \mathbb{R}^M \to \overline{\mathbb{R}}$ with $f(x, y) = f_1(x) + f_2(y)$ the subdifferential satisfies $\partial f(x, y) = \partial f_1(x) \times \partial f_2(y)$ [48, Proposition 10.5].

Finally, the *distance* of $\bar{x} \in \mathbb{R}^N$ to a set $\omega \subset \mathbb{R}^N$ is given by $\operatorname{dist}(\bar{x}, \omega) := \inf_{x\in\omega} \|\bar{x} - x\|$ and we introduce $\|\partial f(\bar{x})\|_- := \inf_{v\in\partial f(\bar{x})} \|v\| = \operatorname{dist}(0, \partial f(\bar{x}))$, which is known as the *lazy slope* of f at $\bar{x}$. Note that $\inf \emptyset := +\infty$ by definition. Furthermore, we have (see [21]) the following lemma.

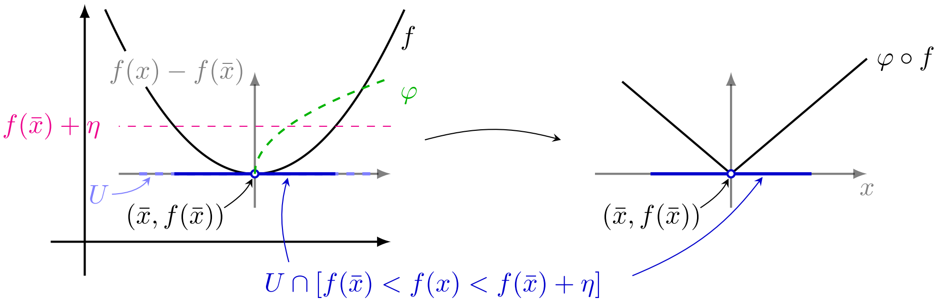

FIG. 1. *Example of the KL property for a smooth function. The composition $\varphi \circ f$ has a slope of magnitude 1 except at $\bar{x}$.*

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

For a function f, we use the notation $[f < \mu] := \{x \in \mathbb{R}^N \,|\, f(x) < \mu\}$. Analogously, we use the same notation for other conditions, for example, $[f \geq \mu]$, $[f = 1]$, etc.

2.2. The Kurdyka–Łojasiewicz property.

DEFINITION 2 (the Kurdyka–Łojasiewicz (KL) property). *Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be an extended real-valued function and let $\bar{x} \in \operatorname{dom} \partial f$. If there exists $\eta \in (0, \infty]$, a neighborhood U of $\bar{x}$, and a continuous concave function $\varphi\colon [0, \eta) \to \mathbb{R}_+$ such that*

$$\varphi(0) = 0, \quad \varphi \in C^1((0, \eta)), \quad \text{and} \quad \varphi'(s) > 0 \text{ for all } s \in (0, \eta),$$

and for all $x \in U \cap [f(\bar{x}) < f(x) < f(\bar{x}) + \eta]$ the Kurdyka–Łojasiewicz inequality

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

holds, then the function has the Kurdyka–Łojasiewicz property at $\bar{x}$.

If, additionally, the function is lower semicontinuous and the property holds for each point in $\operatorname{dom} \partial f$, then f is called a Kurdyka–Łojasiewicz function.

Figure 1, which is taken from [43], shows the idea and the variables appearing in the definition of the KL property for a smooth function. For smooth functions (assume $f(\bar{x}) = 0$), (1) reduces to $\|\nabla(\varphi \circ f)\| \geq 1$ around the point $\bar{x}$, which means that after reparametrization with a *desingularization function* φ the function is sharp. "Since the function φ is used here to turn a singular region—a region in which the gradients are arbitrarily small—into a regular region, i.e., a place where the gradients are bounded away from zero, it is called a desingularization function for f" [5]. It is easy to see that the KL property is satisfied for all nonstationary points [4].

The KL property is satisfied by a large class of functions, namely functions that are definable in an o-minimal structure (see [4, Theorem 14] and [9, Theorem 14]).

THEOREM 3 (nonsmooth KL inequality for definable functions). *Any proper lower semicontinuous function $f\colon X \to \overline{\mathbb{R}}$ that is definable in an* o*-minimal structure $\mathcal{O}$ has the KL property at each point of $\operatorname{dom} \partial f$. Moreover the function φ in Definition 2 is definable in $\mathcal{O}$.*

In particular, semialgebraic and globally subanalytic sets and functions are definable in such a structure. There is even an o-minimal structure that extends the one of globally subanalytic functions with the exponential function (thus the logarithm is also included) [53, 20]. In fact, o-minimal structures can be seen as an axiomatization of the nice properties of semialgebraic functions, and are therefore designed such that the structure is preserved under many operations, for example, pointwise addition and multiplication, composition and inversion. A brief summary of the concepts that are important for this paper can be found in [4].

Before we introduce the general framework and the convergence analysis in the next sections, let us first consider a so-called *uniformization result*, which was proved in [3] for the Łojasiewicz property and adjusted in [12] for the KL property. Its main implication for this paper—like in [12]—is that it allows for a direct proof of the main convergence theorem without the need of an induction argument.

LEMMA 4 (uniformization result [12]). *Let ω be a compact set and let $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper and lower semicontinuous function. Assume that f is constant on ω and satisfies the KL property at each point of ω. Then, there exist $\varepsilon > 0$, $\eta > 0$, and a continuous concave function $\varphi\colon [0, \eta) \to \mathbb{R}_+$ such that*

$$\varphi(0) = 0, \quad \varphi \in C^1((0, \eta)), \quad \textit{and} \quad \varphi'(s) > 0 \textit{ for all } s \in (0, \eta),$$

such that for all $\bar{x} \in \omega$ and all x in the intersection

$$[\operatorname{dist}(x, \omega) < \varepsilon] \cap [f(\bar{x}) < f(x) < f(\bar{x}) + \eta] \tag{2}$$

one has

$$\varphi'(f(x) - f(\bar{x}))\|\partial f(x)\|_- \geq 1.$$

3. An abstract inexact convergence theorem.

In this section, let $\mathcal{F}\colon \mathbb{R}^N \times \mathbb{R}^P \to \overline{\mathbb{R}}$ be a proper lower semicontinuous function that is bounded from below. We analyze convergence of an abstract algorithm that generates a sequence $(x^n)_{n\in\mathbb{N}}$ in $\mathbb{R}^N$ under the following realistic assumptions. Many algorithms, such as the gradient descent method, forward-backward splitting, alternating projection, proximal minimization, heavy-ball method, iPiano, and many more methods, satisfy these assumptions. An application to block coordinate and variable metric iPiano is presented in sections 4 and 5.

Assumption H. Let $(u^n)_{n\in\mathbb{N}}$ be a sequence of parameters in $\mathbb{R}^P$, and let $(\varepsilon_n)_{n\in\mathbb{N}}$ be an ℓ_1-summable sequence of nonnegative real numbers. Moreover, we assume there are sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, and $(d_n)_{n\in\mathbb{N}}$ of nonnegative real numbers, a nonempty finite index set $I \subset \mathbb{Z}$ and $\theta_i \geq 0$, $i \in I$, with $\sum_{i\in I} \theta_i = 1$ such that the following hold.

(H1) *Sufficient decrease condition.* For each $n \in \mathbb{N}$, it holds that

$$\mathcal{F}(x^{n+1}, u^{n+1}) + a_n d_n^2 \leq \mathcal{F}(x^n, u^n).$$

(H2) *Relative error condition.* For each $n \in \mathbb{N}$ (setting $d_j = 0$ for $j \leq 0$), the following holds:

$$b_{n+1}\|\partial\mathcal{F}(x^{n+1}, u^{n+1})\|_- \leq b \sum_{i\in I} \theta_i d_{n+1-i} + \varepsilon_{n+1}.$$

(H3) *Continuity condition.* There exists a subsequence $((x^{n_j}, u^{n_j}))_{j\in\mathbb{N}}$ and $(\tilde{x}, \tilde{u}) \in \mathbb{R}^N \times \mathbb{R}^P$ such that

$$(x^{n_j}, u^{n_j}) \xrightarrow{\mathcal{F}} (\tilde{x}, \tilde{u}) \quad \text{as} \quad j \to \infty.$$

(H4) *Distance condition.* It holds that

$$d_n \to 0 \Longrightarrow \|x^{n+1} - x^n\|_2 \to 0$$

and

$$\exists n' \in \mathbb{N}\colon \forall n \geq n'\colon d_n = 0 \Longrightarrow \exists n'' \in \mathbb{N}\colon \forall n \geq n''\colon x^{n+1} = x^n.$$

(H5) *Parameter condition.* It holds that

$$(b_n)_{n\in\mathbb{N}} \notin \ell_1, \quad \sup_{n\in\mathbb{N}} \frac{1}{b_n a_n} < \infty, \quad \inf_n a_n =: \underline{a} > 0.$$

Let us first discuss how these assumptions generalize previous results and what the perspectives of the newly gained flexibility are. The convergence of the sequence $(x^n)_{n\in\mathbb{N}}$ is proved in Theorem 10.

3.1. Relation to other abstract convergence conditions. The following works explicitly formulate abstract conditions that are used in specific algorithms. Examples of algorithms for which the generalizations are necessary are provided in Appendix A.

Relation to [5]. For a proper lower semicontinuous function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and a sequence $(x^n)_{n\in\mathbb{N}}$, the conditions in [5] are the following:

(ABS13-H1) *For each* $n \in \mathbb{N}$, $f(x^{n+1}) + a\|x^{n+1} - x^n\|_2^2 \leq f(x^n)$.

(ABS13-H2) *For each* $n \in \mathbb{N}$, *there exists* $w^{n+1} \in \partial f(x^{n+1})$ *such that* $\|w^{n+1}\| \leq b\|x^{n+1} - x^n\|_2$.

(ABS13-H3) *There exists a subsequence* $(x^{n_j})_{j\in\mathbb{N}}$ *and* $\tilde{x}$ *such that* $x^{n_j} \to \tilde{x}$ *and* $f(x^{n_j}) \to f(\tilde{x})$ *as* $j \to \infty$.

If the conditions (ABS13-H1)–(ABS13-H3) hold, then Assumption H is also satisfied, which shows that our result is more general. The relation is explicitly shown by setting $\mathcal{F}(x^n, u^n) = f(x^n)$, $u^n = 0$, $a_n = a \in \mathbb{R}$, $b_n = 1$, $I = \{1\}$, $\theta_1 = 1$, $\varepsilon_n = 0$ for all $n \in \mathbb{N}$, and $d_n = \|x^{n+1} - x^n\|_2$.

Relation to [21]. In [21], the conditions in [5] are generalized to a flexible parameter setting and Hilbert spaces. In $\mathbb{R}^N$, the conditions read as follows.

(FGP14-H1) *For each* $n \in \mathbb{N}$, *for some* $a_n > 0$, $f(x^{n+1}) + a_n\|x^{n+1} - x^n\|_2^2 \leq f(x^n)$.

(FGP14-H2) *For each* $n \in \mathbb{N}$, *for some* $b_{n+1} > 0$ *and* $\varepsilon_{n+1} \geq 0$, $b_{n+1}\|\partial f(x^{n+1})\|_- \leq \|x^{n+1} - x^n\|_2 + \varepsilon_{n+1}$.

(FGP14-H3) *The sequences* $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$ *satisfy*

$$a_n \geq \underline{a} > 0 \;\; \textit{for all} \;\; n \in \mathbb{N}, \quad (b_n)_{n\in\mathbb{N}} \notin \ell_1,$$
$$\sup_{n\in\mathbb{N}} \frac{1}{b_n a_n} < \infty, \quad \textit{and} \quad (\varepsilon_n)_{n\in\mathbb{N}} \in \ell_1.$$

The continuity condition (ABS13-H3) is replaced by an f-precompactness assumption. The fact that Assumption H is a generalization of these conditions follows immediately from the relation to [5] and the design of our parameters $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$, in analogy to those in [21]. Our relative error condition (H2) and distance condition (H4) are more general and we allow for a second argument in the objective function u^n whose convergence is not sought in the end, i.e., we allow for a controlled change of the objective function along the iterations.

Relation to [11]. The abstract convergence statement [11, Proposition 4] poses conditions on a triplet of points $\{x^{n-1}, x^n, x^{n+1}\}$ and a function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$. The conditions are the following.[2]

(BP16-H1) *For each* $n \in \mathbb{N}$, $f(x^n) + a\|x^{n+1} - x^n\|_2^2 \le f(x^{n-1})$.

(BP16-H2) *For each* $n \in \mathbb{N}$, $\|\partial f(x^n)\|_- \le b\|x^{n+1} - x^n\|_2$.

(BP16-H3) *There exists a subsequence* $(x^{n_j})_{j\in\mathbb{N}}$ *and* $\tilde{x}$ *such that*

$$x^{n_j} \to \tilde{x} \quad \textit{and} \quad f(x^{n_j}) \to f(\tilde{x}) \quad \textit{as} \quad j \to \infty.$$

In contrast to (ABS13-H2) and (FGP14-H2), the relative error condition (BP16-H2) is explicit (like in [1], or more explicitly discussed in [42, section 2.4]), i.e., x^{n+1} does not appear inside the subdifferential estimate. Setting $d_n = \|x^{n+2} - x^{n+1}\|_2$, $I = \{1\}$, $\theta_1 = 1$, $a_n = a \in \mathbb{R}$, $b_n = 1$, $\varepsilon_n = 0$, $u^n = 0$, and $\mathcal{F}(x^n, u^n) = f(x^n)$ in Assumption H recovers the conditions (BP16-H1)–(BP16-H3). Note that the definition of d_n does not conflict with (H4).

Relation to [44]. The abstract convergence theorem of [44] applies to a sequence $(z^n)_{n\in\mathbb{N}}$ given by $z^n = (x^n, x^{n-1})$ with a sequence $(x^n)_{n\in\mathbb{N}}$ in $\mathbb{R}^N$ for a function $f\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$. The conditions are the following.

(OCBP14-H1) *For each* $n \in \mathbb{N}$, $f(z^{n+1}) + a\|x^n - x^{n-1}\|_2^2 \le f(z^n)$.

(OCBP14-H2) *For each* $n \in \mathbb{N}$, *there exists* $w^{n+1} \in \partial f(z^{n+1})$ *such that*

$$\|w^{n+1}\| \le \frac{b}{2}(\|x^n - x^{n-1}\|_2 + \|x^{n+1} - x^n\|_2).$$

(OCBP14-H3) *There exists a subsequence* $(z^{n_j})_{j\in\mathbb{N}}$ *and* $\tilde{z}$ *such that* $z^{n_j} \to \tilde{z}$ *and* $f(z^{n_j}) \to f(\tilde{z})$ *as* $j \to \infty$.

These conditions are recovered from our framework by setting $\mathcal{F}(z^n, u^n) = f(z^n)$, $d_n = \|x^n - x^{n-1}\|_2$, $a_n = a \in \mathbb{R}$, $b_n = 1$, $I = \{1, 2\}$, $\theta_1 = \theta_2 = \frac{1}{2}$, and $\varepsilon_n = 0$ for all $n \in \mathbb{N}$.

Remark 1. Note that, using the equivalence between norms, the right-hand side of the inequality in (OCBP14-H2) can be bounded from above:

$$\|x^n - x^{n-1}\|_2 + \|x^{n+1} - x^n\|_2 \le \sqrt{2}\|z^{n+1} - z^n\|_2.$$

3.2. Discussion and perspectives. In section 3.1, we have seen that the conditions in Assumption H are more general than previous abstract convergence results. In the following, we provide some discussion, intuition, and perspectives of the conditions in Assumption H.

- Since $\mathcal{F}$ is bounded from below, (H1) requires that $a_n d_n$ tends to 0 as $n \to \infty$. Moreover, as $\inf_n a_n > 0$, this implies that $d_n \to 0$.
- However, $(a_n)_{n\in\mathbb{N}}$ is not a priori assumed to be bounded. The faster a_n tends to ∞, the faster the property $a_n d_n \to 0$ requires d_n to tend to 0.
- If $d_n \to 0$ and, assuming for a moment that $\inf_n b_n > 0$, (H2) implies that $\|\partial\mathcal{F}(x^n, u^n)\|_- \to 0$. However, $(b_n)_{n\in\mathbb{N}}$ may tend to 0, though not too fast because of (H5). The required slow behavior of $b_n \to 0$ will still allow us to conclude that $\liminf_{n\to\infty} \|\partial\mathcal{F}(x^n, u^n)\|_- = 0$.
- The sequence $(d_n)_{n\in\mathbb{N}}$ is introduced as a more general distance measure, which by (H4) is "consistent" with the Euclidean distance. The purpose of this generalization

[2] We neglect the dependence of b in (BP16-H2) on the compact set that contains x^n, as this set will be chosen to be the KL-neighborhood of the set of limit points, which fixes the parameter for sufficiently large n.

is to open the door for Bregman distances [16] without the common assumption of strong convexity or Lipschitz continuity of the gradient. Alternatively, the sequence $(d_n)_{n\in\mathbb{N}}$ can measure the distance between $(x^n)_{n\in\mathbb{N}}$ and a sequence of surrogate points that only asymptotically require $\|x^{n+1}-x^n\|_2 \to 0$. Of course, when distances are only measured with such an abstract distance measure, convergence in the Euclidean sense cannot be expected without further assumptions. A third option, which we explore in this paper, is a sequence $(d_n)_{n\in\mathbb{N}}$ that measures the Euclidean distance only of a block of coordinates of $(x^n)_{n\in\mathbb{N}}$, which leads to block coordinate descent algorithms. A sufficient condition to achieve (H4) is to repeat each block after a finite number of steps (possibly unordered).

• The extension of (H2) to the sum $\sum_{i\in I}\theta_i d_{n+1-i}$ seems to be important for multistep methods such as the heavy-ball method [47], iPiano [44, 43], and other inertial forward-backward splitting methods [15, 36]. For the setting of [36], we provide some details in the appendix.

• The introduction of a sequence $(u^n)_{n\in\mathbb{N}}$ adds some flexibility in the asymptotic behavior of the objective function. For example, in [44], most of the analysis allows for step sizes and other parameters to change in each iteration. However, there is a crucial parameter (δ-parameter inside the Lyapunov function) that is required to be constant for the convergence result. Using the gained flexibility from the sequence $(u^n)_{n\in\mathbb{N}}$, the problem can be resolved. The variable metric iPiano considered in section 4 requires a Lyapunov function that depends on a whole matrix, which, thanks to the sequence $(u^n)_{n\in\mathbb{N}}$ in Assumption H, can change in each iteration (see (18)). Note that this problem occurs due to the definition of the Lyapunov function and does not appear, for example, in [21], where the variable metric is handled in a different way.

3.3. Convergence analysis.

3.3.1. Direct consequences of the descent property. Sufficient decrease (H1) of a certain quantity that can be related to the objective function value is key for the convergence analysis. The following lemma lists a few simple but favorable properties for such sequences.

LEMMA 5. *Let Assumption* H *hold. Then*

(i) $(\mathcal{F}(x^n,u^n))_{n\in\mathbb{N}}$ *is nonincreasing;*

(ii) $(\mathcal{F}(x^n,u^n))_{n\in\mathbb{N}}$ *converges;*

(iii) $\sum_{k=1}^n d_k^2 < +\infty$ *and, therefore,* $d_n \to 0$ *and* $\|x^{n+1}-x^n\|_2 \to 0$ *as* $n\to\infty$.

Proof. Parts (i) and (ii) follow from (H1) and the boundedness from below of $\mathcal{F}$. Part (iii) follows from summing (H1) from $k=1,\dots,n$ and (H4), (H5):

$$\begin{aligned}\underline{a}\sum_{k=1}^n d_k^2 \le \sum_{k=1}^n a_k d_k^2 &\le \sum_{k=1}^n \mathcal{F}(x^k,u^k)-\mathcal{F}(x^{k+1},u^{k+1})\\ &= (\mathcal{F}(x^1,u^1) - \inf_{(x,u)\in\mathbb{R}^N\times\mathbb{R}^P}\mathcal{F}(x,u)) < +\infty.\end{aligned}$$ □

3.3.2. Direct consequences for the set of limit points. Like in [12], we can verify some results about the set of limit points (that depends on a certain initialization) of a bounded sequence $((x^n,u^n))_{n\in\mathbb{N}}$, which is given by

$$\omega(x^0,u^0) := \limsup_{n\to\infty}\{(x^n,u^n)\}.$$

This definition uses the outer set-limit of a sequence of singletons, which is the same as the set of cluster points in a different notation. Moreover, we denote by $\omega_{\mathcal{F}}(x^0,u^0)$

the subset of limit points that is generated along $\mathcal{F}$-attentive subsequences, i.e.,

$$\omega_{\mathcal{F}}(x^0,u^0) := \big\{(\bar{x},\bar{u}) \in \omega(x^0,u^0) \,|\, (x^{n_j},u^{n_j}) \xrightarrow{\mathcal{F}} (\bar{x},\bar{u}) \text{ for } j\to\infty\big\}.$$

We collect a few results that are of independent interest.

LEMMA 6. *Let Assumption* H *hold and let* $((x^n,u^n))_{n\in\mathbb{N}}$ *be a bounded sequence.*

(i) *The set* $\omega_{\mathcal{F}}(x^0,u^0)$ *is nonempty and the set* $\omega(x^0,u^0)$ *is nonempty and compact.*

(ii) $\mathcal{F}$ *is constant and finite on* $\omega_{\mathcal{F}}(x^0,u^0)$.

Proof.

(i) By (H3), there exists a subsequence $((x^{n_j},u^{n_j}))_{j\in\mathbb{N}}$ of $((x^n,u^n))_{n\in\mathbb{N}}$ that converges to $(\tilde{x},\tilde{u})$, where at the same time the function values along this subsequence converge to $\mathcal{F}(\tilde{x},\tilde{u})$, and therefore $\lim_{j\to\infty}(x^{n_j},u^{n_j}) \in \omega_{\mathcal{F}}(x^0,u^0)$ and $\omega_{\mathcal{F}}(x^0,u^0)$ is nonempty. The nonemptiness of $\omega(x^0,u^0)$ is clear and the compactness of $\omega(x^0,u^0)$ is a direct consequence of its definition as an outer set-limit and the boundedness of $((x^n,u^n))_{n\in\mathbb{N}}$.

(ii) By Lemma 5(ii) $(\mathcal{F}(x^n,u^n))_{n\in\mathbb{N}}$ converges to some $\tilde{\mathcal{F}}\in\mathbb{R}$. For any $(\bar{x},\bar{u})\in \omega_{\mathcal{F}}(x^0,u^0)$ there exists a subsequence $((x^{n_j},u^{n_j}))_{j\in\mathbb{N}}$ that $\mathcal{F}$-converges to $(\bar{x},\bar{u})$, and therefore

$$\tilde{\mathcal{F}} = \lim_{j\to\infty}\mathcal{F}(x^{n_j},u^{n_j}) = \mathcal{F}(\bar{x},\bar{u}),$$

which shows that $\mathcal{F}$ is constant on $\omega_{\mathcal{F}}(x^0,u^0)$. ∎

LEMMA 7. *Let Assumption* H *hold and* $((x^n,u^n))_{n\in\mathbb{N}}$ *be a bounded sequence. Denote by* $\Pi_x(\omega) = \{x\in\mathbb{R}^N \,|\, (x,u)\in\omega\}$ *the projection of* $\omega\in\mathbb{R}^N\times\mathbb{R}^P$ *onto the first* N *coordinates. Then we have the following results:*

(i) *The set* $\Pi_x(\omega(x^0,u^0))$ *is connected.*

(ii) *If* $(u^n)_{n\in\mathbb{N}}$ *converges, then the set* $\omega(x^0,u^0)$ *is connected.*

(iii) *It holds that*

$$\lim_{n\to\infty}\operatorname{dist}((x^n,u^n),\omega(x^0,u^0)) = 0.$$

Proof. Part (i) is a simple application of the connectedness results [12, Lemma 5] and the fact that $\|x^{n+1}-x^n\|_2\to 0$ for $n\to\infty$ by Lemma 5(iii). Part (ii) follows in almost the same manner, as convergence of u^n implies $\|u^{n+1}-u^n\|_2\to 0$ as $n\to\infty$. Part (iii) is a direct consequence of the definition of the set of limit points. ∎

LEMMA 8. *Let Assumption* H *hold, let* $((x^n,u^n))_{n\in\mathbb{N}}$ *be a bounded sequence, and let* $\sum_{n=0}^\infty d_n<\infty$. *Then, the set* $\omega_{\mathcal{F}}(x^0,u^0)\subset \operatorname{crit}\mathcal{F}$.

Proof. Let $(\bar{x},\bar{u})\in\omega(x^0,u^0)$. Then, since $(b_n)_{n\in\mathbb{N}}\notin\ell_1$ holds, from (H2), we have $(\varepsilon_n)_{n\in\mathbb{N}}\in\ell_1$ and

$$\sum_{n=0}^\infty b_n\|\partial\mathcal{F}(x^n,u^n)\|_- \le b\sum_{n=0}^\infty\sum_{i\in I}\theta_i d_{n-i} + \sum_{n=0}^\infty\varepsilon_n<\infty,$$

which implies $\liminf_{n\to\infty}\|\partial\mathcal{F}(x^n,u^n)\|_- = 0$. For $(\bar{x},\bar{u})\in\omega_{\mathcal{F}}(x^0,u^0)$ the subsequence $((x^{n_j},u^{n_j}))_{j\in\mathbb{N}}$ $\mathcal{F}$-converges to $(\bar{x},\bar{u})$ as $j\to\infty$ and Lemma 1 implies that $0\in\partial\mathcal{F}(\bar{x},\bar{u})$, which was to be proved. ∎

COROLLARY 9. *Let Assumption* H *hold and let* $((x^n, u^n))_{n\in\mathbb{N}}$ *be a bounded sequence. Suppose* $\mathcal{F}$ *is continuous on the set* $W \cap \operatorname{dom}\mathcal{F}$ *with an open set* $W \supset \omega(x^0, u^0)$ *(e.g.,* $\mathcal{F}$ *is continuous on* $\operatorname{dom}\mathcal{F}$*). Then*

$$\omega(x^0, u^0) = \omega_{\mathcal{F}}(x^0, u^0).$$

Proof. Let

$$(x^{n_j}, u^{n_j}) \to (\bar{x}, \bar{u}) \in \omega(x^0, u^0) \quad \text{as} \quad j \to \infty.$$

There is a neighborhood $V \subset W$ with $(\bar{x}, \bar{u}) \in V$ such that $(x^{n_j}, u^{n_j}) \in V \cap \operatorname{dom}\mathcal{F}$ for sufficiently large $j \in \mathbb{N}$, and continuity of $\mathcal{F}$ implies $(x^{n_j}, u^{n_j}) \xrightarrow{\mathcal{F}} (\bar{x}, \bar{u})$, and thus $\omega(x^0, u^0) \subset \omega_{\mathcal{F}}(x^0, u^0)$. The converse inclusion holds by definition. ∎

3.3.3. The convergence theorem.

THEOREM 10. *Suppose* $\mathcal{F}$ *is a proper lower semicontinuous Kurdyka–Łojasiewicz function that is bounded from below. Let* $(x^n)_{n\in\mathbb{N}}$ *be a bounded sequence generated by an abstract algorithm parametrized by a bounded sequence* $(u^n)_{n\in\mathbb{N}}$ *that satisfies Assumption* H. *Assume that* $\mathcal{F}$*-attentive convergence holds along converging subsequences of* $((x^n, u^n))_{n\in\mathbb{N}}$*, i.e.,* $\omega(x^0, u^0) = \omega_{\mathcal{F}}(x^0, u^0)$*. Then, the following hold.*

(i) *The sequence* $(d_n)_{n\in\mathbb{N}}$ *satisfies*

$$\sum_{k=0}^{\infty} d_k < +\infty, \tag{3}$$

i.e., the trajectory of the sequence $(x^n)_{n\in\mathbb{N}}$ *has finite length with respect to the abstract distance measures* $(d_n)_{n\in\mathbb{N}}$.

(ii) *Suppose* d_k *satisfies* $\|x^{k+1} - x^k\|_2 \le \bar{c} d_{k+k'}$ *for some* $k' \in \mathbb{Z}$ *and* $\bar{c} \in \mathbb{R}$*. Then*

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|_2 < +\infty, \tag{4}$$

and the trajectory of the sequence $(x^n)_{n\in\mathbb{N}}$ *has a finite Euclidean length, and thus* $(x^n)_{n\in\mathbb{N}}$ *converges to* $\tilde{x}$ *from* (H3).

(iii) *Furthermore, if* $(u^n)_{n\in\mathbb{N}}$ *is a converging sequence, then each limit point of* $((x^n, u^n))_{n\in\mathbb{N}}$ *is a critical point, which in the situation of* (ii) *is the unique point* $(\tilde{x}, \tilde{u})$ *from* (H3).

Proof. (The proof follows the same technique as [12].) By (H3) there exists a subsequence $((x^{n_j}, u^{n_j}))_{j\in\mathbb{N}}$ such that $(x^{n_j}, u^{n_j}) \xrightarrow{\mathcal{F}} (\tilde{x}, \tilde{u})$ as $j \to \infty$. If there is n' such that $\mathcal{F}(x^{n'}, u^{n'}) = \mathcal{F}(\tilde{x}, \tilde{u})$, then (H1) implies that $\mathcal{F}(x^n, u^n) = \mathcal{F}(\tilde{x}, \tilde{u})$ for all $n \ge n'$, thus also $a_n d_n^2 = 0$ and by $\underline{a} > 0$ (see (H4)) $d_n = 0$ for all $n \ge n'$. Therefore, (H4) shows that $x^{n+1} = x^n$ for all $n \ge n''$ for some $n'' \in \mathbb{N}$, and by induction $(x^n)_{n\in\mathbb{N}}$ becomes stationary (i.e., $x^n = x^{n''}$ for all $n \ge n''$) and the statement is obvious.

Now, we can assume that $\mathcal{F}(x^n, u^n) > \mathcal{F}(\tilde{x}, \tilde{u})$ for all $n \in \mathbb{N}$. Moreover, nonincreasingness of $(\mathcal{F}(x^n, u^n))_{n\in\mathbb{N}}$ by (H1) implies that for all $\eta > 0$ there exists $n_1 \in \mathbb{N}$ such that $\mathcal{F}(\tilde{x}, \tilde{u}) < \mathcal{F}(x^n, u^n) < \mathcal{F}(\tilde{x}, \tilde{n}) + \eta$ for all $n \ge n_1$. By definition there is also a region of attraction for the sequence $(x^n, u^n)_{n\in\mathbb{N}}$, i.e., for all $\varepsilon > 0$ there exists $n_2 \in \mathbb{N}$ such that $\operatorname{dist}((x^n, u^n), \omega(x^0, u^0)) < \varepsilon$ holds for all $n \ge n_2$. In total, we know that for all $n \ge n_0 := \max\{n_1, n_2\}$ the sequence $((x^n, u^n))_{n\in\mathbb{N}}$ lies in the set

$$[\mathcal{F}(\tilde{x}, \tilde{u}) < \mathcal{F}(x, u) < \mathcal{F}(\tilde{x}, \tilde{u}) + \eta] \cap [\operatorname{dist}((x, u), \omega(x^0, u^0)) < \varepsilon].$$

Combining the facts that $\omega(x^0,u^0)=\omega_{\mathcal{F}}(x^0,u^0)$ is nonempty and compact from Lemma 6(i) with $\mathcal{F}$ being finite and constant on $\omega(x^0,u^0)$ from Lemma 6(ii) allows us to apply Lemma 4 with $\omega=\omega(x^0,u^0)$. Therefore, there are φ, η, ε as in Lemma 4 such that for $n>n_0$,

$$\varphi'(\mathcal{F}(x^n,u^n)-\mathcal{F}(\tilde{x},\tilde{u}))\|\partial\mathcal{F}(x^n,u^n)\|_-\geq 1 \tag{5}$$

holds on ω. Plugging (H2) into (5) yields

$$\varphi'(\mathcal{F}(x^n,u^n)-\mathcal{F}(\tilde{x},\tilde{u}))\geq b_n\left(b\sum_{i\in I}\theta_i d_{n-i}+\varepsilon_n\right)^{-1}. \tag{6}$$

By concavity of φ, letting $m>n$,

$$\begin{aligned}D^{\varphi}_{n,m}&:=\varphi(\mathcal{F}(x^n,u^n)-\mathcal{F}(\tilde{x},\tilde{u}))-\varphi(\mathcal{F}(x^m,u^m)-\mathcal{F}(\tilde{x},\tilde{u}))\\&\geq\varphi'(\mathcal{F}(x^n,u^n)-\mathcal{F}(\tilde{x},\tilde{u}))(\mathcal{F}(x^n,u^n)-\mathcal{F}(x^m,u^m)),\end{aligned}$$

and using (6) and (H1) we infer

$$D^{\varphi}_{n,n+1}\geq\frac{b_na_nd_n^2}{b\sum_{i\in I}\theta_id_{n-i}+\varepsilon_n}\quad\Leftrightarrow\quad d_n^2\leq\left(\sum_{i\in I}\theta'_id_{n-i}+\varepsilon'_n\right)\left(\frac{b'}{a_nb_n}D^{\varphi}_{n,n+1}\right),$$

where we use the substitutions

$$\overline{\theta}:=\sum_{j\in I}\theta_j,\quad b':=b\overline{\theta},\quad \theta'_i:=\theta_i/\overline{\theta},\quad\text{and}\quad\varepsilon'_n:=\varepsilon_n/b'.$$

Applying $2\sqrt{\alpha\beta}\leq\alpha+\beta$ for all $\alpha,\beta\geq 0$, we obtain (set $c:=\sup_n\frac{b'}{a_nb_n}<\infty$ (by (H4)))

$$2d_n\leq\frac{b'}{a_nb_n}D^{\varphi}_{n,n+1}+\sum_{i\in I}\theta'_id_{n-i}+\varepsilon'_n\leq cD^{\varphi}_{n,n+1}+\sum_{i\in I}\theta'_id_{n-i}+\varepsilon'_n.$$

Now summing this inequality over $k=n_0,\dots,n$ yields

$$2\sum_{k=n_0}^n d_k\leq\sum_{k=n_0}^n\sum_{i\in I}\theta'_id_{k-i}+c\sum_{k=n_0}^nD^{\varphi}_{k,k+1}+\sum_{k=n_0}^n\varepsilon'_k. \tag{7}$$

The first sum on the right-hand side can be rewritten (using the substitution $j=k-i$) as follows:[3]

$$\sum_{k=n_0}^n\sum_{i\in I}\theta'_id_{k-i}=\sum_{i\in I}\sum_{j=n_0-i}^{n-i}\theta'_id_j=\left(\sum_{i\in I}\theta'_i\right)\sum_{j=n_0}^nd_j+\sum_{i\in I}\sum_{j=n_0-i}^{n_0-1}\theta'_id_j+\sum_{i\in I}\sum_{j=n+1}^{n-i}\theta'_id_j.$$

Using $\sum_{i\in I}\theta'_i=1$ and rearranging terms in (7) yields

$$\sum_{k=n_0}^nd_k\leq\sum_{i\in I}\sum_{j=n_0-i}^{n_0-1}\theta'_id_j+\sum_{i\in I}\sum_{j=n+1}^{n-i}\theta'_id_j+c\sum_{k=n_0}^nD^{\varphi}_{k,k+1}+\sum_{k=n_0}^n\varepsilon'_k.$$

[3]We use the convention that the summation is zero when the start index is larger than the termination index.

From this inequality, we conclude that $\lim_{n\to\infty}\sum_{k=0}^{n} d_k < +\infty$. The first and second terms of the right-hand side are finite summations and $d_n \to 0$ as $n \to \infty$. The third term equals $cD^{\varphi}_{n_0,n+1}$, which is bounded from above by $\varphi(\mathcal{F}^{n_0}(x^{n_0}) - \mathcal{F}(\tilde{x})) < +\infty$. The last term is finite by assumption $(\varepsilon_n)_{n\in\mathbb{N}} \in \ell_1$, which, in total, verifies (i).

Part (ii) is a consequence of (i) and the fact that for arbitrary $m > n > 0$,

$$\|x^m - x^n\|_2 \le \sum_{k=n}^{m-1} \|x^{k+1} - x^k\|_2 \le c \sum_{k=n}^{m-1} d_{k+k'} < +\infty$$

holds, which shows that $(x^n)_{n\in\mathbb{N}}$ is a Cauchy sequence (the right-hand side vanishes for $n, m \to \infty$). Therefore, $x^n \to \tilde{x}$ as $n \to \infty$, which verifies (ii). Using (i) and (ii), (iii) is a direct consequence of Lemma 8. ∎

4. Variable metric iPiano.

We consider a structured nonsmooth nonconvex optimization problem with a proper lower semicontinuous extended-valued function $h\colon \mathbb{R}^N \to \overline{\mathbb{R}}$, $N \ge 1$, that is bounded from below by some value $\underline{h} > -\infty$:

$$\min_{x\in\mathbb{R}^N} h(x), \quad h(x) = f(x) + g(x). \tag{8}$$

The function $f\colon \mathbb{R}^N \to \mathbb{R}$ is assumed to be C^1-smooth (possibly nonconvex) with L-Lipschitz continuous gradient on $\operatorname{dom} g$, $L > 0$. Further, let the function $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be simple (possibly nonsmooth and nonconvex) and prox-bounded, i.e., there exists $\lambda > 0$ such that

$$e_\lambda g(x) := \inf_{y\in\mathbb{R}^N} g(y) + \frac{1}{2\lambda}\|y - x\|^2 > -\infty$$

for some $x \in \mathbb{R}^N$. Saying "g is simple" refers to the fact that the associated proximal map can be solved efficiently for the global optimum.

We propose Algorithm 1 to find a critical point $x^* \in \operatorname{dom}\partial g$ of h, which in this case is characterized by

$$-\nabla f(x^*) \in \partial g(x^*),$$

where ∂g denotes the limiting subdifferential. The parameter restrictions are discussed in Lemma 11 and Remark 3.

Depending on the properties of g, the step-size parameter α_n and the inertial parameter β_n must satisfy different conditions. We analyse the properties when g is convex, semiconvex, or nonconvex in a concise manner. If g is semiconvex with respect to the metric induced by $A \in \mathbb{S}_{++}(N)$, let m be the semiconvexity parameter, i.e., $m \in \mathbb{R}$ is the largest value such that $g(x) - \frac{m}{2}\|x\|_A^2$ is convex. For convex functions $m = 0$ and for strongly convex functions $m > 0$. Instead of considering the situation where g is nonconvex as a semiconvex function with "$m = -\infty$", we introduce a "flag variable" $\sigma \in \{0, 1\}$, which is 1 if g is semiconvex and 0 if g is nonconvex. Note that if $\sigma = 1$, the property of semiconvexity is satisfied for any $A \in \mathbb{S}_{++}(N)$, but with possibly changing modulus. Therefore, sometimes the metric is not explicitly specified.

Algorithm 1. *Variable metric inertial proximal algorithm for nonconvex optimization (vmiPiano).*

- ***Parameter***.
  - *Let* $(\alpha_n)_{n\in\mathbb{N}}$ *be a sequence of positive step-size parameters.*
  - *Let* $(\beta_n)_{n\in\mathbb{N}}$ *be a sequence of nonnegative parameters.*
  - *Let* $(A_n)_{n\in\mathbb{N}}$ *be a sequence of matrices* $A_n \in \mathbb{S}_{++}(N)$ *such that* $A_n \preceq \mathrm{id}$ *and* $\inf_n \varsigma(A_n) > 0$.
  - *Let* $\sigma = 1$ *if* g *is semiconvex and* $\sigma = 0$ *otherwise.*
- ***Initialization***. *Choose a starting point* $x^0 \in \operatorname{dom} h$ *and set* $x^{-1} = x^0$.
- ***Iterations*** $(\boldsymbol{n \geq 0})$. *Update:*

$$
\begin{aligned}
y^n &= x^n + \beta_n(x^n - x^{n-1}),\\
x^{n+1} &\in \arg\min_{x\in\mathbb{R}^N} \; Q^n(x; x^n),\\
Q^n(x;x^n) &:= g(x) + \langle \nabla f(x^n), x - x^n\rangle + \frac{1}{2\alpha_n}\|x - y^n\|^2_{A_n},
\end{aligned} \tag{9}
$$

where $L_n > \sigma m_n$ *is determined such that*

$$
f(x^{n+1}) \leq f(x^n) + \left\langle \nabla f(x^n), x^{n+1} - x^n \right\rangle + \frac{L_n}{2}\|x^{n+1} - x^n\|^2_{A_n} \tag{10}
$$

holds and $\alpha_n,\ \beta_n$ *with* $\inf_n \alpha_n > 0$ *are chosen such that (see, e.g., Lemma* 11*)*

$$
\delta_n^\sigma := \frac{1}{2}\left(\frac{1+\sigma-\beta_n}{\alpha_n} - (L_n - \sigma m_n)\right) \quad \textit{and} \quad \gamma_n := \delta_n^\sigma - \frac{\beta_n}{2\alpha_n} \tag{11}
$$

satisfy

$$
\inf_n \gamma_n > 0 \quad \textit{and} \quad \delta_n^\sigma \|x^n - x^{n-1}\|^2_{A_n} \leq \delta_{n-1}^\sigma \|x^n - x^{n-1}\|^2_{A_{n-1}}, \tag{12}
$$

where $m_n \in \mathbb{R}$ *denotes the semiconvexity modulus of* g *w.r.t.* $A_n \in \mathbb{S}_{++}(N)$ *(if* $\sigma = 1$*).*

Remark 2. The minimization problem in (9) is equivalent to (constant terms are dropped)

$$
\arg\min_{x\in\mathbb{R}^N} \; g(x) + \langle \nabla f(x^n), x - x^n\rangle - \frac{\beta_n}{\alpha_n}\left\langle x^n - x^{n-1}, x - x^n\right\rangle_{A_n} + \frac{1}{2\alpha_n}\|x - x^n\|^2_{A_n}. \tag{13}
$$

A necessary condition for x^{n+1} is

$$
x \in (\mathrm{id} + \alpha_n A_n^{-1}\partial g)^{-1}\left(x^n - \alpha_n A_n^{-1}\nabla f(x^n) + \beta_n(x^n - x^{n-1})\right). \tag{14}
$$

For a convex function g, inverting the expression $\mathrm{id} + \alpha_n A_n^{-1}\partial g$ yields a unique solution and the inclusion can be replaced by an equality. Here, the operator is set-valued.

4.1. Parameter selection in Algorithm 1.

Lemma 11. *A necessary condition for the sequences* $(\alpha_n)_{n\in\mathbb{N}}$ *and* $(\beta_n)_{n\in\mathbb{N}}$ *to satisfy* $\gamma_n \geq c > 0$ *for all* $n \in \mathbb{N}$ *is*

$$
\alpha_n \leq \frac{1+\sigma-2\beta_n}{L_n - \sigma m_n + 2c} \quad \textit{and} \quad \beta_n < \frac{1+\sigma}{2}.
$$

Proof. The bounds directly follow from $\inf_n \gamma_n > 0$. ∎

A simple but restrictive choice of parameters, whose proof is obvious, is the following.

LEMMA 12. *The following conditions are sufficient to satisfy* (12)*:*

$$\delta_{n+1}^\sigma \le \delta_n^\sigma, \quad A_{n+1} \preceq A_n, \quad \textit{and } \alpha_n,\ \beta_n \textit{ according to Lemma } 11.$$

For a flexible choice of the metric, the following lemma constructs a rule for selecting feasible parameters.

LEMMA 13. *A sufficient condition to satisfy* (12)*, which allows for a flexible choice of* $(A_n)_{n\in\mathbb{N}}$*, is the following: suppose* $x^n \neq x^{n-1}$*. Given* δ_{n-1}^σ *and* A_{n-1}*, select* A_n *and define*

$$\begin{aligned}&\tilde\delta_n^\sigma = \frac{\delta_{n-1}^\sigma \|x^n - x^{n-1}\|_{A_{n-1}}^2}{\|x^n - x^{n-1}\|_{A_n}^2}, \quad \alpha_n = \frac{1+\sigma-2\beta_n}{L_n - \sigma m_n + 2c},\\ &\textit{and} \quad \beta_n = \frac{1+\sigma}{2}\frac{b-1}{b-\frac12} \quad \textit{with} \quad b = \frac{\tilde\delta_n^\sigma + (L_n - \sigma m_n)}{L_n - \sigma m_n + 2c}.\end{aligned} \tag{15}$$

Proof. The conditions are derived by $\delta_n^\sigma = \frac12(\frac{1+\sigma-\beta_n}{\alpha_n} - (L_n - \sigma m_n)) \le \tilde\delta_n^\sigma$, which is equivalent to

$$\alpha_n \ge \frac{1+\sigma-\beta_n}{\tilde\delta_n^\sigma + (L_n - \sigma m_n)} =: \underline{\alpha}_n.$$

The condition in Lemma 11 is

$$\alpha_n \le \frac{1+\sigma-2\beta_n}{L_n - \sigma m_n + 2c} =: \overline{\alpha}_n,$$

which allows for the choice of α_n in (15), since β_n is selected such that $\overline{\alpha}_n - \underline{\alpha}_n \ge 0$. ∎

Note that, if $x^n = x^{n-1}$, then $y^n = x^n$ and (9) is a proximal gradient step without momentum, which allows for any metric A_n and L_n such that (10) holds; set $\alpha_n = \frac{1}{L_n}$.

Remark 3.

- The assumption in (10) is satisfied, for example, if f has an L-Lipschitz continuous gradient with $A_n = \mathrm{id}$, or when a local estimate of the Lipschitz constant L_n is known (again $A_n = \mathrm{id}$).
- The additional hyperparameters δ_n^σ and γ_n can be seen as an disadvantage, however, actually they allow for a constructive selection of the step-size parameters (see Lemma 13). For example, in [15] such hyperparameters do not appear and only existence of parameters that satisfy certain conditions can be guaranteed.
- Unlike in [44, 43], where the sequence δ_n is assumed to be stationary after a finite number of iterations to obtain the final convergence result, here the restrictions for δ_n and A_n are very loose: essentially boundedness is required.

4.2. Convergence analysis of Algorithm 1. As mentioned before, we want to take advantage of g being semiconvex. The next lemmas are essential for that.

LEMMA 14. *Let g be proper semiconvex with modulus $m \in \mathbb{R}$ with respect to the metric induced by $A \in \mathbb{S}_{++}(N)$. Then, for any $\bar x \in \operatorname{dom}\partial g$ it holds that*

$$g(x) \ge g(\bar x) + \langle \bar v, x - \bar x\rangle + \frac{m}{2}\|x - \bar x\|_A^2 \quad \forall x \in \mathbb{R}^N \textit{ and } \bar v \in \partial g(\bar x).$$

Proof. Fix $\tilde{x} \in \operatorname{dom} g$ and apply the subgradient inequality to $g_m(x) := g(x) - \frac{m}{2}\|x - \tilde{x}\|_A^2$ around the point $\bar{x}$, i.e., it holds that

$$g_m(x) \geq g_m(\bar{x}) + \langle \bar{w}, x - \bar{x} \rangle \quad \forall x \in \mathbb{R}^N \text{ and } \bar{w} \in \partial g_m(\bar{x}).$$

Note that $\bar{w}$ is an element from the (convex) subdifferential. Due to the smoothness of $\frac{m}{2}\|x - \tilde{x}\|_A^2$, we can use the summation rule for the limiting subdifferential to obtain

$$\partial g_m(\bar{x}) = \partial \left(g - \frac{m}{2}\| \cdot - \tilde{x}\|_A^2 \right)(\bar{x}) = \partial g(\bar{x}) - mA(\bar{x} - \tilde{x}),$$

and therefore, replacing $\bar{w}$ by $\bar{v} - mA(\bar{x} - \tilde{x})$ with $\bar{v} \in \partial g(\bar{x})$ in the subgradient inequality above, we obtain after using

$$2 \langle \bar{x} - \tilde{x}, x - \bar{x} \rangle_A = \|x - \tilde{x}\|_A^2 - \|\bar{x} - \tilde{x}\|_A^2 - \|x - \bar{x}\|_A^2$$

that the inequality

$$g_m(x) + \frac{m}{2}\|x - \tilde{x}\|_A^2 \geq g_m(\bar{x}) + \frac{m}{2}\|\bar{x} - \tilde{x}\|_A^2 + \frac{m}{2}\|x - \bar{x}\|_A^2 + \langle \bar{v}, x - \bar{x} \rangle \qquad \forall x \in \mathbb{R}^N \text{ and } \bar{v} \in \partial g(\bar{x})$$

holds, which implies the statement. □

LEMMA 15. *Let $\sigma = 1$ if g is proper semiconvex with modulus $m \in \mathbb{R}$ with respect to the metric induced by $A \in \mathbb{S}_{++}(N)$ and $\sigma = 0$ otherwise. Then it holds that*

$$Q^n(x^{n+1}; x^n) + \frac{\sigma}{2}\left(m + \frac{1}{\alpha_n} \right) \|x^{n+1} - x^n\|_A^2 \leq Q^n(x^n; x^n). \tag{16}$$

Proof. If $\sigma = 1$, then the function $x \mapsto Q^n(x; x^n)$ from (9) is semiconvex with modulus $m + \frac{1}{\alpha_n}$. We apply Lemma 14 with $x = x^n$, $\bar{x} = x^{n+1}$, and $0 \in \partial Q^n(x^{n+1}, x^n)$ to deduce (16) with $\sigma = 1$. If $\sigma = 0$, then we simply observe that x^{n+1} is a minimizer of (9), which yields (16) with $\sigma = 0$. □

Verification of Assumption H. We define the proper lower semicontinuous function

$$\begin{aligned} &\mathcal{F} \colon \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N \times N} \times \mathbb{R} \to \overline{\mathbb{R}} \\ &\qquad \text{given by} \quad \mathcal{F}(x, y, A, \delta) := H_{(\delta, A)}(x, y) := h(x) + \delta \|x - y\|_A^2 \end{aligned} \tag{17}$$

for some $A \in \mathbb{S}_{++}(N)$ and $\delta \in \mathbb{R}$. Regarding the variables in Assumption H, the u-component of $\mathcal{F}$ is treated as $u = (A, \delta)$, which allows the function $\mathcal{F}$ to change depending on the metric A and another parameter δ. Convergence will be derived for the x and y variables only.

The following proposition verifies (H1), with $d_n = \|x^n - x^{n-1}\|_2$ and $a_n = \gamma_n$.

PROPOSITION 16 (descent property). *Let the variables and parameters be given as in Algorithm 1. Then, it holds that*

$$H_{(\delta_n^\sigma, A_n)}(x^{n+1}, x^n) \leq H_{(\delta_n^\sigma, A_n)}(x^n, x^{n-1}) - \gamma_n \varsigma(A_n) \|x^n - x^{n-1}\|_2^2, \tag{18}$$

and the sequence $(H_{(\delta_n^\sigma, A_n)}(x^n, x^{n-1}))_{n \in \mathbb{N}}$ is monotonically decreasing, which verifies condition (H1) *with $\mathcal{F}$ as in* (17)*, $d_n = \|x^n - x^{n-1}\|_2$, and $a_n = \gamma_n \varsigma(A_n)$.*

Proof. Combining (9) (in the equivalent form (13)) with (10) and (16) yields

$$
\begin{aligned}
&f(x^{n+1}) + g(x^{n+1}) + \frac{\sigma}{2}\left(m + \frac{1}{\alpha_n}\right)\|x^{n+1} - x^n\|_{A_n}^2 \\
&\le f(x^n) + \left\langle \nabla f(x^n), x^{n+1} - x^n \right\rangle + \frac{L_n}{2}\|x^{n+1} - x^n\|_{A_n}^2 + g(x^n) \\
&\quad - \left\langle \nabla f(x^n), x^{n+1} - x^n \right\rangle + \frac{\beta_n}{\alpha_n}\left\langle x^{n+1} - x^n, x^n - x^{n-1}\right\rangle_{A_n} - \frac{1}{2\alpha_n}\|x^{n+1} - x^n\|_{A_n}^2 \\
&= f(x^n) + g(x^n) + \frac{\beta_n}{\alpha_n}\left\langle x^{n+1} - x^n, x^n - x^{n-1}\right\rangle_{A_n} + \left(\frac{L_n}{2} - \frac{1}{2\alpha_n}\right)\|x^{n+1} - x^n\|_{A_n}^2,
\end{aligned}
$$

and using $\langle a, b\rangle_M \le \frac{1}{2}(\|a\|_M^2 + \|b\|_M^2)$ for any $a, b \in \mathbb{R}^N$ and $M \in \mathbb{S}_{++}(N)$ implies the inequality

$$
h(x^{n+1}) \le h(x^n) + \frac{\beta_n}{2\alpha_n}\|x^n - x^{n-1}\|_{A_n}^2 - \frac{1}{2}\left(\frac{1 + \sigma - \beta_n}{\alpha_n} - (L_n - \sigma m)\right)\|x^{n+1} - x^n\|_{A_n}^2.
$$

Finally, rearranging terms yields

$$
\begin{aligned}
&h(x^{n+1}) + \delta_n^\sigma\|x^{n+1} - x^n\|_{A_n}^2 \\
&\qquad \le h(x^n) + \delta_n^\sigma\|x^n - x^{n-1}\|_{A_n}^2 - \left(\delta_n^\sigma - \frac{\beta_n}{2\alpha_n}\right)\|x^n - x^{n-1}\|_{A_n}^2. \qquad \square
\end{aligned}
$$

The parametrization of the step sizes is chosen as in [43] (see [43, Lemma 6.3] for well-definedness of the parameters). Therefore, we obtain the same step-size restrictions here, but with the flexibility to change the metric in each iteration.

Remark 4. The proof shows that instead of (13) we could also consider

$$
\arg\min_{x \in \mathbb{R}^N} \; g(x) + \langle \nabla f(x^n), x - x^n\rangle - \frac{\beta_n}{\alpha_n}\left\langle x^n - x^{n-1}, x - x^n\right\rangle + \frac{1}{2\alpha_n}\|x - x^n\|_{A_n}^2, \tag{19}
$$

which yields a slightly different algorithm, but step-size restrictions are the same. This expression differs from (13) in the metric of the inner product with coefficient β_n/α_n.

Next, we prove the relative error condition (condition (H2)) with $b_n \equiv 1$ and $\varepsilon_n \equiv 0$, $I = \{1, 2\}$, and $\theta_1 = \theta_2 = \frac{1}{2}$. We derive a bound on the (limiting) subgradient of the function h, which we employ to bound the subgradient of the function $\mathcal{F}$.

LEMMA 17. *Let the variables and parameters be given as in Algorithm* 1*. Then, there exists* $b > 0$ *such that*

$$
\|\partial h(x^{n+1})\|_- \le \frac{b}{2}\left(\|x^{n+1} - x^n\|_2 + \|x^n - x^{n-1}\|_2\right).
$$

Proof. Equation (14) can be used to specify an element from $\partial g(x^{n+1})$, namely

$$
A_n\frac{x^n - x^{n+1}}{\alpha_n} - \nabla f(x^n) + \frac{\beta_n}{\alpha_n}A_n(x^n - x^{n-1}) \in \partial g(x^{n+1}),
$$

which implies

$$
\begin{aligned}
\|\partial h(x^{n+1})\|_- &= \|\nabla f(x^{n+1}) + \partial g(x^{n+1})\|_- \\
&\le \left(\frac{\|A_n\|}{\alpha_n} + L\right)\|x^{n+1} - x^n\|_2 + \frac{\beta_n}{\alpha_n}\|A_n\|\|x^n - x^{n-1}\|_2.
\end{aligned}
$$

Using the Lipschitz continuity of ∇f and $A \preceq \mathrm{id}$, the statement is verified. $\square$

Proposition 18. *Let the variables and parameters be given as in Algorithm* 1. *Then, there exists $b > 0$ such that*

$$\|\partial\mathcal{F}(x^{n+1}, x^n, A_{n+1}, \delta^\sigma_{n+1})\|_- \leq \frac{b}{2}\left(\|x^{n+1} - x^n\|_2 + \|x^n - x^{n-1}\|_2\right),$$

which verifies condition (H2) *with $\mathcal{F}$ as in* (17), $d_n = \|x^n - x^{n-1}\|_2$, $b_n \equiv 1$, $I = \{1, 2\}$, $\theta_1 = \theta_2 = \frac{1}{2}$, *and* $\varepsilon_n \equiv 0$.

Proof. Thanks to the summation rule of the limiting subdifferential for the sum of $(x, y, A, \delta) \mapsto h(x)$ and the smooth function $(x, y, A, \delta) \mapsto \delta\|x^{n+1} - x^n\|^2_A$, we can compute the limiting subdifferential by estimating the partial derivatives. We obtain

$$\begin{aligned} \partial_x\mathcal{F}(x, y, A, \delta) &= \partial h(x) + 2\delta A(x - y), \\ \partial_y\mathcal{F}(x, y, A, \delta) &= \nabla_y\mathcal{F}(x, y, A, \delta) = -2A\delta A(x - y), \end{aligned} \tag{20}$$

$$\begin{aligned} \partial_A\mathcal{F}(x, y, A, \delta) &= \nabla_A\mathcal{F}(x, y, A, \delta) = \delta(x - y) \otimes (x - y), \\ \partial_\delta\mathcal{F}(x, y, A, \delta) &= \nabla_\delta\mathcal{F}(x, y, A, \delta) = \|x - y\|^2_A. \end{aligned} \tag{21}$$

In order to verify (H2), let $\mathcal{F}^{n+1} := \mathcal{F}(x^{n+1}, x^n, A_{n+1}, \delta^\sigma_{n+1})$ and we use

$$\|w^{n+1}\|_2 \leq \|w^{n+1}_x\|_2 + \|w^{n+1}_y\|_2 + \|w^{n+1}_A\|_2 + \|w^{n+1}_\delta\|_2,$$

where $w^{n+1} \in \partial\mathcal{F}^{n+1}$ with block coordinates $w^{n+1}_x \in \partial_x\mathcal{F}^{n+1}$, $w^{n+1}_y = \nabla_y\mathcal{F}^{n+1}$, $w^{n+1}_A = \nabla_A\mathcal{F}^{n+1}$, and $w^{n+1}_\delta = \nabla_\delta\mathcal{F}^{n+1}$. We obtain the relative error bound (H2) using Lemma 17, $A_{n+1} \preceq \mathrm{id}$, boundedness of δ^σ_{n+1}, and the fact that for a sequence $r_n \to 0$ for some $n_0 \in \mathbb{N}$ it holds that $r_n^2 \leq r_n$ for all $n \geq n_0$. In detail, we use

$$\begin{aligned} \|w^{n+1}_A\|_2 &\leq \delta^\sigma_{n+1}\sum_{i,j}|x^{n+1}_i - x^n_i| \cdot |x^{n+1}_j - x^n_j| \leq c\sum_{i,j}|x^{n+1}_j - x^n_j| \\ &\leq cc'\sum_i\|x^{n+1} - x^n\|_2 \leq cc'c''\|x^{n+1} - x^n\|_2, \end{aligned}$$

where c is the maximal (over the coordinates i) bound for the converging sequences $|x^{n+1}_i - x^n_i| \to 0$ as $n \to \infty$, the dimensionally dependent constant $c' = \sqrt{N}$ provides the norm equivalence of $\|\cdot\|_1$ and $\|\cdot\|_2$, and $c'' = N$ simplifies the summation. ∎

The next proposition shows that converging subsequences of the sequence generated by Algorithm 1 always $\mathcal{F}$-converge to the limit point, i.e., $\omega(x^0, u^0) = \omega_{\mathcal{F}}(x^0, u^0)$ is automatically satisfied, which implies (H3) when the algorithm generates a bounded sequence.

Proposition 19. *Let the variables and parameters be given as in Algorithm* 1. *Then, any convergent subsequence* $((x^{n_j+1}, x^{n_j}, A_{n_j}, \delta^\sigma_{n_j}))_{j\in\mathbb{N}}$ *actually $\mathcal{F}$-converges to a point* $(x^*, x^*, A_*, \delta^\sigma_*)$, *which verifies* (H3) *for a bounded sequence* $((x^n, u^n))_{n\in\mathbb{N}}$ *with $\mathcal{F}$ as in* (17).

Proof. Let $(x^{n_j+1}, x^{n_j}, A_{n_j}, \delta^\sigma_{n_j})$ be a subsequence converging to some $(x^*, x^*, A_*, \delta^\sigma_*)$. The continuity statement follows ($Q^n(x^{n+1}; x^n) \leq Q^n(x; x^n)$ for all $x \in \mathbb{R}^N$ from (9)) from

$$\begin{aligned} g(x^{n_j+1}) + \left\langle\nabla f(x^{n_j}), x^{n_j+1} - x^{n_j}\right\rangle + \frac{1}{2\alpha_{n_j}}\|x^{n_j+1} - y^{n_j}\|^2_{A_{n_j}} \\ \leq g(x^*) + \langle\nabla f(x^{n_j}), x^* - x^{n_j}\rangle + \frac{1}{2\alpha_{n_j}}\|x^* - y^{n_j}\|^2_{A_{n_j}}. \end{aligned}$$

Due to Lemma 5(iii), $\|x^{n_j+1} - x^{n_j}\| \to 0$, and hence $\|y^{n_j} - x^{n_j}\| \to 0$, which shows that $y^{n_j} \to x^*$ as $j \to \infty$. Moreover, since f is continuously differentiable, $\nabla f(x^{n_j})$ converges as $j \to \infty$, and hence it is bounded. Therefore, considering the limit superior of $j \to \infty$ of both sides of the inequality shows that $\limsup_{j\to\infty} g(x^{n_j+1}) \le g(x^*)$, which combined with the lower semicontinuity of g implies $\lim_{j\to\infty} g(x^{n_j+1}) = g(x^*)$, and thus the statement follows, since f is continuously differentiable. ∎

Using the results that we just derived, we can prove convergence of the variable metric iPiano method (Algorithm 1) to a critical point. Unlike the abstract convergence theorems in [5, 21, 44], the finite length property is derived for the coordinates from a subspace only, which allows for a lot of flexibility. Critical points are characterized in the proof of Proposition 18 (see (20)), where zero in the partial subdifferential (actually the partial derivative) with respect to y, A, or δ implies $x = y$ without imposing conditions on the δ- or A-coordinate. Thus, we have

$$0 \in \partial\mathcal{F}(x, y, A, \delta) \Leftrightarrow \Big(0 \in \partial h(x) \times 0_y \times 0_A \times 0_\delta \text{ and } x = y\Big) \Leftrightarrow \Big(0 \in \partial h(x) \text{ and } x = y\Big),$$

where we indicate the size of the zero variables by the respective coordinate variable. As a consequence, $0 \in \mathcal{F}(x^*, x^*, \delta, A) \Leftrightarrow 0 \in \partial h(x^*)$. These considerations lead to the following convergence theorem.

THEOREM 20. *Suppose $\mathcal{F}$ in* (17)*,* (8) *is a proper lower semicontinuous Kurdyka–Lojasiewicz function that is bounded from below. Let $(x^n)_{n\in\mathbb{N}}$ be generated by Algorithm* 1 *and bounded with valid variables and parameters as in the description of this algorithm. Then, the sequence $(x^n)_{n\in\mathbb{N}}$ satisfies*

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|_2 < +\infty, \tag{22}$$

and $(x^n)_{n\in\mathbb{N}}$ converges to a critical point of (8).

Proof. Verify the condition in Assumption H and apply Theorem 10. Set $d_n = \|x^n - x^{n-1}\|_2$, $a_n = \gamma_n \varsigma(A_n)$, $b_n \equiv 1$, $\varepsilon_n \equiv 0$, $I = \{1, 2\}$, $\theta_1 = \theta_2 = \frac{1}{2}$; then (H1), (H2), and (H3) are proved in Propositions 16, 18, and 19, and (H4), (H5) are immediate from the bounds on the parameters. ∎

Remark 5. Thanks to [8, 9] the KL property holds for proper lower semicontinuous functions that are definable in an o-minimal structure, e.g., semialgebraic functions. Since o-minimal structures are stable under various operations, $\mathcal{F}$ is a KL function if h is definable in an o-minimal structure. Therefore, Theorem 20 can be applied to, for instance, a proper lower semicontinuous semialgebraic function h in (8).

5. Block coordinate variable metric iPiano.

We consider a structured nonsmooth nonconvex optimization problem with a proper lower semicontinuous extended-valued function $h: \mathbb{R}^N \to \overline{\mathbb{R}}$, $N \ge 1$, that is bounded from below by some value $\underline{h} > -\infty$:

$$\min_{\mathbf{x}\in\mathbb{R}^N} h(\mathbf{x}), \quad h(\mathbf{x}) := f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_J) + \sum_{i=1}^{J} g_i(\mathbf{x}_i), \tag{23}$$

where the N dimensions are partitioned into J blocks of (possibly different dimensions) $(N_1, \dots, N_J)$, i.e., $\mathbf{x} \in \mathbb{R}^N$ can be decomposed as $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_J)$. The function $f: \mathbb{R}^N \to \mathbb{R}$ is assumed to be block C^1-smooth (possibly nonconvex) with block Lipschitz continuous gradient on $\operatorname{dom} g_1 \times \operatorname{dom} g_2 \times \cdots \times \operatorname{dom} g_J$, i.e., $\mathbf{x}_i \mapsto \nabla_{\mathbf{x}_i} f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_J)$ is Lipschitz continuous. Further, let the function $g_i : \mathbb{R}^{N_i} \to \overline{\mathbb{R}}$ be simple (possibly nonsmooth and nonconvex) and prox-bounded.

Working with block algorithms can be simplified with some appropriate notation, which we introduce now. We denote by $\mathbf{x}_{\bar{i}} := (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_J)$ the vector containing all blocks but the ith one.

Algorithm 2 is a straightforward extension of Algorithm 1 to problems of class (23) with a block coordinate structure. In each iteration, the algorithm applies one iteration of iPiano to the problem restricted to a certain block. The formulation of the algorithm allows blocks to be updated in an almost arbitrary order. In the end, the only restriction is that each block must be updated infinitely often, which is a more flexible rule than in [46]. The discussion of the parameters from section 4.1 extends straightforwardly to the block coordinate structure.

We seek a critical point $\mathbf{x}^* \in \operatorname{dom} \partial g$ of h, which in this case is characterized by

$$-\nabla f(\mathbf{x}^*) \in \partial g_1(\mathbf{x}_1^*) \times \partial g_2(\mathbf{x}_2^*) \times \cdots \times \partial g_J(\mathbf{x}_J^*).$$

In fact if we apply Algorithm 2 to (8) from the preceding section (i.e., $J = 1$), we recover the variable metric iPiano algorithm (Algorithm 1). For $\beta_{n,i} = 0$ for all $n \in \mathbb{N}$ and $i \in \{1, \ldots, J\}$, the algorithm is known as the block coordinate variable metric forward-backward (BC-VMFB) algorithm [19]. If, additionally $A_{n,i} = \mathrm{id}$ for all n and i, the algorithm is referred to as proximal alternating linearized minimization (PALM) [12]. An inertial block coordinate version (without variable metric) is proposed in [46] as iPALM.

Verification of Assumption H. In order to prove convergence of this algorithm, we can make use of the results of the preceding section for the variable metric iPiano algorithm. We consider a function

$$\mathcal{F} : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N_1 \times N_1} \times \cdots \times \mathbb{R}^{N_J \times N_J} \times \mathbb{R}^J \to \overline{\mathbb{R}} \tag{24}$$

given by (set $\mathbf{A} := (A_1, \ldots, A_J)$, $A_i \in \mathbb{R}^{N_i \times N_i}$, $\mathbf{\Delta} := (\delta_1, \ldots, \delta_J)$)

$$\mathcal{F}(\mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{\Delta}) = H_{\mathbf{\Delta}, \mathbf{A}}(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) + \sum_{i=1}^{J} \delta_i \|\mathbf{x}_i - \mathbf{y}_i\|_{A_i}^2 .$$

THEOREM 21. *Suppose $\mathcal{F}$ in* (24), (23) *is a proper lower semicontinuous KL function (e.g., h is semialgebraic; cf. Remark* 5*) that is bounded from below. Let $(\mathbf{x}^n)_{n\in\mathbb{N}}$ be generated by Algorithm* 2 *and bounded with valid variables and parameters as in the description of this algorithm. Assume that each block coordinate is updated after a finite number of $n' \in \mathbb{N}$ steps. Then, the sequence $(\mathbf{x}^n)_{n\in\mathbb{N}}$ satisfies*

$$\sum_{k=0}^{\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 < +\infty, \tag{25}$$

and $(\mathbf{x}^n)_{n\in\mathbb{N}}$ converges to a critical point of (23).

Algorithm 2. *Block coordinate variable metric iPiano.*

- ***Parameter***. *For all* $i \in \{1, \dots, J\}$,
  - *let* $(\alpha_{n,i})_{n\in\mathbb{N}}$ *be a sequence of positive step-size parameters;*
  - *let* $(\beta_{n,i})_{n\in\mathbb{N}}$ *be a sequence of nonnegative parameters;*
  - *let* $(A_{n,i})_{n\in\mathbb{N}}$ *be a sequence of matrices* $A_{n,i} \in \mathbb{S}_{++}(N_i)$ *such that* $A_{n,i} \preceq \mathrm{id}$ *and* $\inf_{n,i} \varsigma(A_{n,i}) > 0$*;*
  - *let* $\sigma_i = 1$ *if* g_i *is semiconvex and* $\sigma_i = 0$ *otherwise.*
- ***Initialization***. *Choose a starting point* $x^0 \in \operatorname{dom} h$ *and set* $x^{-1} = x^0$.
- ***Iterations*** $(\boldsymbol{n \geq 0})$. *Update: select* $j_n \in \{1, \dots, J\}$ *and compute*

(26)
$$
\begin{aligned}
\mathbf{y}_{j_n}^n &= \mathbf{x}_{j_n}^n + \beta_{n,j_n}(\mathbf{x}_{j_n}^n - \mathbf{x}_{j_n}^{n-1}),\\
\mathbf{x}_{j_n}^{n+1} &\in \arg\min_{x\in\mathbb{R}^{N_{j_n}}} \; Q^{j_n}(x; \mathbf{x}_{j_n}^n),\\
&\qquad Q^{j_n}(x; \mathbf{x}_{j_n}^n) := g_{j_n}(x) + \left\langle \nabla_{\mathbf{x}_{j_n}} f(\mathbf{x}^n), x - \mathbf{x}_{j_n}^n \right\rangle + \frac{1}{2\alpha_{n,j_n}} \|x - \mathbf{y}_{j_n}^n\|_{A_{n,j_n}}^2,\\
\mathbf{x}_{\bar{j}_n}^{n+1} &= \mathbf{x}_{\bar{j}_n}^n,\\
\mathbf{x}_{\bar{j}_n}^n &= \mathbf{x}_{\bar{j}_n}^{n-1},
\end{aligned}
$$

where $L_n > \sigma m_n$ *is determined such that*

$$
f(\mathbf{x}^{n+1}) \leq f(\mathbf{x}^n) + \left\langle \nabla_{\mathbf{x}_{j_n}} f(\mathbf{x}^n), \mathbf{x}_{j_n}^{n+1} - \mathbf{x}_{j_n}^n \right\rangle + \frac{L_n}{2} \|\mathbf{x}_{j_n}^{n+1} - \mathbf{x}_{j_n}^n\|_{A_{n,j_n}}^2 \tag{27}
$$

holds and α_{n,j_n}, β_{n,j_n} *with* $\inf_{n,j} \alpha_{n,j} > 0$ *are chosen such that*

(28)
$$
\delta_{n,j_n}^{\sigma_{j_n}} := \frac{1}{2}\left(\frac{1 + \sigma_{j_n} - \beta_{n,j_n}}{\alpha_{n,j_n}} - (L_n - \sigma_{j_n} m_n)\right) \quad \text{and} \quad \gamma_{n,j_n} := \delta_{n,j_n}^{\sigma_{j_n}} - \frac{\beta_{n,j_n}}{2\alpha_{n,j_n}}
$$

satisfy

$$
\inf_{n,j} \gamma_{n,j} > 0 \quad \text{and} \quad \delta_{n+1,j_n}^{\sigma_{j_n}} \|\mathbf{x}_{j_n}^{n+1} - \mathbf{x}_{j_n}^n\|_{A_{n+1,j_n}}^2 \leq \delta_{n,j_n}^{\sigma_{j_n}} \|\mathbf{x}_{j_n}^{n+1} - \mathbf{x}_{j_n}^n\|_{A_{n,j_n}}^2, \tag{29}
$$

where $m_n \in \mathbb{R}$ *denotes the semiconvexity modulus of* g_{j_n} *w.r.t.* $A_{j_n} \in \mathbb{S}_{++}(N_{j_n})$ *(if* $\sigma_{j_n} = 1$*).*

Set $A_{n+1,\bar{j}_n} = A_{n,\bar{j}_n}$, $\delta_{n+1,\bar{j}_n}^{\sigma_{j_n}} = \delta_{n,\bar{j}_n}^{\sigma_{j_n}}$.

Proof. As the nth iteration of Algorithm 2 reads exactly the same as in Algorithm 1 but applied to the block coordinate j_n only, we can directly apply Proposition 16 and obtain

$$
H_{(\boldsymbol{\Delta}_n^\sigma, \mathbf{A}_n)}(\mathbf{x}^{n+1}, \mathbf{x}^n) \leq H_{(\delta_n^\sigma, A_n)}(\mathbf{x}^n, \mathbf{x}^{n-1}) - \gamma_{n,j_n} \varsigma(A_{n,j_n}) \|\mathbf{x}_{j_n}^n - \mathbf{x}_{n_j}^{n-1}\|_2^2,
$$

and the function H is monotonically decreasing along the iterations, i.e., the parameters in the algorithm are chosen such that one step on an arbitrary block decreases the value of H unless the block coordinate is already stationary.

Since the nonsmooth part of the optimization problem (23) is additively separated, the estimation of the subdifferential is easy as it reduces to the Cartesian product of the subdifferential with respect to each block. Therefore, Proposition 18 can be used analogously to deduce

$$\|\partial\mathcal{F}(\mathbf{x}^{n+1},\mathbf{y}^{n+1},\mathbf{A}_{n+1},\mathbf{\Delta}_{n+1})\|_{-} \le \frac{b}{2}\left(\|\mathbf{x}_{j_n}^{n+1}-\mathbf{x}_{j_n}^{n}\|_2^2+\|\mathbf{x}_{j_n}^{n}-\mathbf{x}_{j_n}^{n-1}\|_2^2\right).$$

Under the assumption that each block is updated at least after n' iterations, the continuity results from Proposition 19 can be transferred easily to the setting of Algorithm 2, i.e., we can conclude that any convergent subsequence of block coordinates actually $\mathcal{F}$-converges to the limit point ($\lim_{k\to\infty} g_i(\mathbf{x}_i^{n_k}) = g_i(\mathbf{x}_i^*)$ for each block $i \in \{1,\ldots,J\}$ and f is continuous anyway).

Therefore, the conditions in Assumption H are verified by $d_n = \|\mathbf{x}_{j_n}^{n} - \mathbf{x}_{j_n}^{n-1}\|_2$, $a_n = \gamma_{n,j_n}\varsigma(A_{n,j_n})$, $u_n = (\mathbf{\Delta}_n^\sigma, \mathbf{A}_n)$, $b_n \equiv 1$, $\varepsilon_n \equiv 0$, $I = \{1,2\}$, and $\theta_1 = \theta_2 = \frac{1}{2}$. Condition (H4) is also satisfied because of the finite repetition of the updates, and (H5) is clearly satisfied. ∎

6. Numerical application.

6.1. A Mumford–Shah-like problem. The continuous Mumford–Shah problem is given formally by

$$\min_{w,\Gamma} \frac{\lambda}{2}\int_\Omega |w-I|^2\,dx + \int_{\Omega\setminus\Gamma} |\nabla w|^2\,dx + \gamma|\Gamma|, \tag{30}$$

where $w\colon \Omega \to \mathbb{R}$ is an image on the image domain $\Omega \subset \mathbb{R}^2$ and $I\colon \Omega\to\mathbb{R}$ is a given noisy image; $|\Gamma|$ measures the length of the jump set Γ. Intuitively, a solution w must be smooth except on a possible jump set Γ, and approximate I. The positive parameters λ and γ steer the importance of each term. In order to solve the problem, the jump set Γ needs to be represented by a mathematical object that is amenable to a numerical implementation.

Therefore, we consider the well-known Ambrosio–Tortorelli approximation [2] given by

$$\min_{w,z} \frac{\lambda}{2}\int_\Omega |w-I|^2\,dx + \int_\Omega z^2|\nabla w|^2\,dx + \gamma\int_\Omega \varepsilon|\nabla z|^2 + \frac{(z-1)^2}{4\varepsilon}\,dx, \tag{31}$$

where $\varepsilon > 0$ is a fixed parameter and $z\colon\Omega\to[0,1]$ is a (soft) edge indicator function, also called a phase-field. The last integral is shown to Gamma-converge to the length of the jump set of (30) as $\varepsilon\to 0$.

In this section, we solve a slight variation of this problem. Instead of an image denoising model we are interested in an inpainting problem (as shown in Figure 2), which is usually more difficult. In image inpainting, the true information about the original image is only given on a subset $[c=1]$ of the image domain (black pixels in Figure 2(b)), where $c\colon\Omega\to\{0,1\}$—the original image I is unknown on $[c=0]$ (white part, Figure 2(b)). In [22], the idea of image inpainting is pushed to a limit and used for PDE-based image compression, i.e., the inpainting mask $[c=1]$ is a small subset of Ω. Usually a simple PDE is used for reconstructing the original image based on its gray values given only on mask points, for instance, linear diffusion in [39] (result given in Figure 2(d)). When the inpainting mask is optimized, linear-diffusion-based inpainting is shown to be competitive with JPEG and sometimes with JPEG2000. Therefore, using a more general inpainting model combined with an optimized inpainting mask is expected to improve this performance. We consider the

(a) *Original image I.* (b) *Mask c (90% unknown).* (c) *Inpainting w using* (32). (d) *Linear diffusion inpainting.*

FIG. 2. *Example for image inpainting/compression. The gray values of the original image* (a) *are stored only at the mask points* (b)*, where known values are black* $[c = 1]$ *and unknown ones are white* $[c = 0]$*. Based on* 10% *known gray values, the original image is reconstructed in* (c) *with the Ambrosio–Tortorelli inpainting* (32) *that we evaluate algorithmically in this paper, and in* (d) *with a simple linear diffusion model* [39] *that arises as a special case of* (32) *when the edge set z is fixed to* 1 *everywhere on the image domain* Ω.

model

$$\begin{aligned} &\min_{w,z} \int_\Omega z^2 |\nabla w|^2 \, dx + \gamma \int_\Omega \varepsilon |\nabla z|^2 + \frac{(z-1)^2}{4\varepsilon} \, dx \\ &\text{s.t. } w(x) = I(x) \quad \forall x \in [c = 1], \end{aligned} \tag{32}$$

which extends the linear diffusion model by optimizing for an additional edge set z. The linear diffusion model is recovered when fixing $z = 1$ on Ω. Since we want to evaluate our algorithms, we neglect the development made for finding an optimal inpainting mask and generate the mask by randomly selecting 10% as known pixels.

From now on, we discretize the problem, and do so with a slight abuse of notation. We use the same symbols to denote the discrete counterparts of the above introduced variables: $I \in \mathbb{R}^N$ is the (vectorized[4]) original image, $c \in \mathbb{R}^N$ is the (inpainting) mask, $w \in \mathbb{R}^N$ is the optimization variable (representing a vectorized image), and $z \in [0, 1]^N$ represents the jump (or edge) set of (30). The continuous gradient ∇ is replaced by a discrete derivative operator $D \in \mathbb{R}^{2N \times N}$ that implements forward differences in horizontal $D_1 \in \mathbb{R}^{N \times N}$ and vertical $D_2 \in \mathbb{R}^{N \times N}$ directions with homogeneous boundary conditions, i.e., forward differences across the image boundary are set to 0. Our discretized model of (32) reads

$$\begin{aligned} &\min_{w,z} \frac{1}{2} \| \operatorname{diag}(z)(D_1 w) \|_2^2 + \frac{1}{2} \| \operatorname{diag}(z)(D_2 w) \|_2^2 + \frac{\gamma\varepsilon}{2} \| Dz \|_2^2 + \frac{\gamma}{4\varepsilon} \| z - 1 \|_2^2 \\ &\text{s.t. } w_i = I_i \quad \forall i \in \{1, \dots, N\} \text{ with } c_i = 1, \end{aligned} \tag{33}$$

where $\operatorname{diag}\colon \mathbb{R}^N \to \mathbb{R}^{N \times N}$ puts a vector on the diagonal of a matrix. Figure 4 shows the input data, the reconstructed image, and the reconstructed edge set for $\varepsilon = 0.1$ and $\gamma = 1/400$, and the number of pixels $N = 551 \cdot 414 = 228114$.

[4]The columns of the image are stacked to form a long vector.

In the following, we evaluate several algorithms that use a variable metric. Let

$$g_1(w) := \delta_X(w) \text{ with } X := \{w \in \mathbb{R}^N \,|\, w_i = I_i \text{ if } c_i = 1\}, \qquad g_2(z) := \frac{\gamma}{4\varepsilon}\|z-1\|_2^2,$$
$$f(w,z) := \frac{1}{2}\left(\|\operatorname{diag}(z)(D_1 w)\|_2^2 + \|\operatorname{diag}(z)(D_2 w)\|_2^2 + \gamma\varepsilon\|Dz\|_2^2\right).$$

We can apply iPiano to (8) with $x = (w,z)$ and $g(x) = (g_1(w), g_2(z))$, or block coordinate iPiano to (23) with $\mathbf{x}_1 = w$ and $\mathbf{x}_2 = z$.

In order to determine a suitable metric, we first compute the derivatives of f,

$$\begin{aligned}\nabla_w f(w,z) &= \left(D_1^\top \operatorname{diag}(z^2) D_1 + D_2^\top \operatorname{diag}(z^2) D_2\right) w,\\ \nabla_z f(w,z) &= \left(\operatorname{diag}((D_1 w)^2) + \operatorname{diag}((D_2 w)^2) + \gamma\varepsilon D^\top D\right) z,\end{aligned}$$

where the squares are understood to be coordinatewise. A feasible metric for block coordinate variable metric iPiano (`BC-VM-iPiano`) must satisfy (27). Therefore, for the w-update step (z is fixed), we require $A_{n,w}$ (the metric w.r.t. the block of w coordinates) to satisfy

$$\langle \nabla_w f(w,z) - \nabla_w f(w',z) - A_{n,w}(w-w'), w-w'\rangle \le 0$$

for all $w,\ w'$, which is achieved, for example, by a diagonal matrix $A_{n,w}$ given by

$$(A_{n,w})_{i,i} = \sum_{j=1}^{N} \left| \left(D_1^\top \operatorname{diag}(z^2) D_1 + D_2^\top \operatorname{diag}(z^2) D_2\right)_{i,j} \right| \tag{34}$$

for all $i \in \{1,\dots,N\}$. In order to avoid numerical problems, we add a small numerical constant 10^{-9} to the diagonal of $A_{n,w}$. For the z-update (w is fixed), analogously, we require $A_{n,z}$ (the metric w.r.t. the block of z coordinates) to satisfy

$$\langle \nabla_w f(w,z) - \nabla_w f(w,z') - A_{n,z}(z-z'), z-z'\rangle \le 0$$

for all $z,\ z'$, which is achieved, for example, by a diagonal matrix $A_{n,z}$ given by

$$(A_{n,z})_{i,i} = \sum_{j=1}^{N} \left| \left(\operatorname{diag}((D_1 w)^2) + \operatorname{diag}((D_2 w)^2) + \gamma\varepsilon D^\top D\right)_{i,j} \right| \tag{35}$$

for all $i \in \{1,\dots,N\}$. Note that compared to (27) the metric contains the scaling $L_{n,w}$ and $L_{n,z}$, respectively. For constant step-size schemes ($A_{n,w} = A_{n,z} = \mathrm{id}$) we use $L_w \le 8$ and[5] $L_z \le 2 + 8\gamma\varepsilon$.

Besides `BC-VM-iPiano`, we test forward-backward splitting (`FB`) with constant step-size scheme $\alpha = 2/\max(L_w, L_z)$, block coordinate forward-backward splitting (`BC-FB`) with step sizes $\alpha_w = 2/L_w$ and $\alpha_z = 2/L_z$ (this method is also known as PALM [12]), variable metric forward-backward splitting (`VM-FB`) with the metric (34) and (35) as a composed diagonal matrix, block coordinate variable metric forward-backward splitting (`BC-VM-FB`) with the metric (34) and (35), iPiano (`iPiano`) with constant step-size scheme $\alpha = 2(1-\beta)/\max(L_w, L_z)$, block coordinate iPiano (`BC-iPiano`) with constant step-size scheme $\alpha_w = 2(1-\beta)/L_w$ and $\alpha_z = 2(1-\beta)/L_z$,

[5] Note that I is normalized to $[0,1]$ and, thus, we observed that w stays in $[0,1]$ too. Therefore $(D_1 w)_i^2$ is in $[0,1]$.

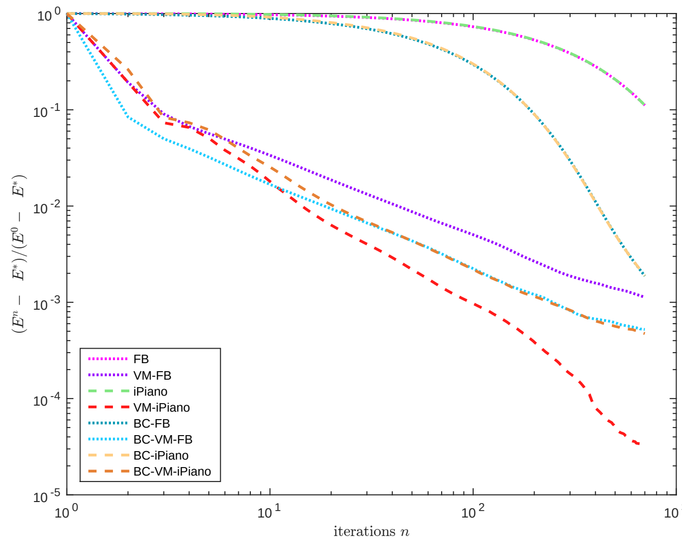

FIG. 3. *Number of iterations vs. relative objective value for solving* (33)*. The performance is significantly improved for methods that take a variable metric into account. Intuitively, this means that the coordinates of the optimization variable are irregularly scaled along the iterations. The variable metric version of iPiano shows the best performance.*

variable metric iPiano (`VM-iPiano`) with the metric (34) and (35) as a composed diagonal matrix, and block coordinate variable metric iPiano (`BC-VM-iPiano`) with the metric (34) and (35). For all methods that incorporate an inertial parameter, it is set to $\beta = 0.7$. The step size for the variable metric algorithms is $\alpha = 2(1-\beta)$ times the inverse (block coordinate) metric (cf. (14)).

Note that `FB`, `VM-FB`, `iPiano`, and `VM-iPiano` are heuristic approaches as the smooth part of the objective is not globally Lipschitz continuous in both variables at the same time. The metric that is used for `VM-FB` and `VM-iPiano` is actually not feasible, as (34) and (35) are not sufficient to guarantee that the metric induces a quadratic majorizer to the function f (cf. (10)). The gradient is not linear with respect to both coordinates. The gradient is linear only if one coordinate is fixed. Nevertheless, in our practical experiments, the methods converged. In future work, we want to analyze if this inaccuracy can be compensated for by making use of relative error conditions, which are not yet incorporated into the algorithms.

We solve problem (33) with all methods up to 1000 iterations and define E^* as the minimal objective value that is achieved among all methods. Let E^0 be the initial value. Figure 3 plots the decrease of the relative objective value $(E^n - E^*)/(E^0 - E^*)$ along the iterations n on a logarithmic scale on both axes (see online version for color figures).

The performance of `FB` and `iPiano` are nearly identical as they do not explore the different scaling of w- and z-coordinates, unlike `BC-FB` and `BC-iPiano`. As both block coordinates seem to "live" on a different scale, block coordinate methods are favorable. However, as the immense performance speed-up of the variable metric methods shows, the irregular scaling also happens to be present among different w-coordinates and z-coordinates. Throughout the experiments, we have noticed that

(a) *Input to model* (33). (b) *Inpainting w using* (33). (c) *Edges z using* (33).

FIG. 4. *Solution to problem* (33). (a) *shows the inpainting mask from Figure* 2(b) *weighted with the gray values from Figure* 2(a). (b) *shows the solution image w and* (c) *the solution edge set z of* (33). *Although the model is nonconvex, visually all algorithms resulted in similar solutions. Figure* 3 *shows that the final objective values differ.*

for optimization problems where regularization (like smoothness between pixels) is important, inertial methods seem to perform slightly better in general. For this experiment, variable metric iPiano shows the best performance and sets the value for E^*, the lowest objective value among all methods after 1000 iterations. Note that the computational cost per iteration is nearly the same for all methods.

7. Conclusion. In this paper, we presented a convergence analysis for abstract inexact generalized descent methods based on the KL-inequality that unifies and generalizes the analysis in Attouch, Bolte, and Svaiter [5], Frankel, Garrigos, and Peypouquet [21], Ochs et al. [44], Bolte and Pauwels [11], and several other more explicit algorithms. The novel convergence theorem allows for more flexibility in the design of algorithms. In more detail, algorithms that imply a descent on a proper lower semicontinuous parametric function and satisfy a certain flexible relative error condition are considered. The parametric function can be seen as an objective function that may vary along the iterations. The gained flexibility is used to formulate a variable metric version of iPiano (an inertial forward-backward splitting-like method). Moreover, thanks to usage of a generic distance measure in the abstract convergence theorem, we obtain a block coordinate variable metric version of iPiano almost for free. Finally, the algorithms are shown to perform well on the practical problem of image compression using a Mumford–Shah-like regularization.

As future work, we will investigate whether the gained flexibility can be used, for example, to prove the convergence of (inertial) Bregman proximal descent methods with Bregman functions that are not required to be strongly convex or to have a Lipschitz continuous gradient.

Appendix A. Relation to algorithms with analogue convergence guarantees.

In recent works, the convergence analysis of algorithms for nonsmooth nonconvex optimization problems often follows along the lines of the proof methodology suggested in [12], i.e., the convergence is explicitly verified, although it suffices to verify the abstract conditions in [5]. In the following, for several such algorithms, the relation to the abstract conditions in [5, 21, 44] and Assumption H is shown. For [32, 33, 36], the generalizations of our paper are necessary to cast them in the abstract framework.

Note that we do not provide an exhaustive list of examples. Most of the algorithms mentioned in the introduction fall into our unifying abstract setting.

Relation to PALM [12]. In [12], the general proof methodology is introduced. Thanks to a uniformization result of the KL-inequality, which we also use in this paper (see Lemma 4), the convergence proof was simplified compared to [5]. Lemma 3(i) of [12] verifies (ABS13-H1), Lemma 4 of [12] shows (ABS13-H2), and Lemma 5(i) of [12] contains the continuity statement (ABS13-H3).

Relation to [15]. An inertial algorithm for the sum of two nonconvex functions was proposed in this paper. The setting is slightly more general than [44] as the nonsmooth part of the objective is allowed to be nonconvex. The proximal subproblems are formulated with respect to Bregman distances that are required to be strongly convex and with Lipschitz continuous gradient, which provides a lower and upper bound in the Euclidean metric for the Bregman distance terms. The proof of convergence is, hence, analogous to that of [44]. However, unlike in [44], the sufficient decrease condition uses $d_n = \|x^{n+1} - x^n\|_2$ instead of $\|x^n - x^{n-1}\|_2$. Both conditions obviously fall into the more general set of conditions in Assumption H. The conditions in Assumption H are verified in [15, (H1)–(H3), Page 13] in analogy with (OCBP14-H1)–(OCBP14-H3) for which we provide the details in section 3.1.

Relation to [32]. A Douglas–Rachford splitting algorithm for solving nonsmooth nonconvex problems of the form

$$\min_{x\in\mathbb{R}^N} f(x) + g(x), \tag{36}$$

where f has Lipschitz continuous gradient and g is proper lower semicontinuous, is proposed. The algorithm generates sequences $(x^n)_{n\in\mathbb{N}}$, $(y^n)_{n\in\mathbb{N}}$, and $(z^n)_{n\in\mathbb{N}}$ according to the following update scheme ($\gamma > 0$):

$$\begin{aligned} y^{n+1} &\in \operatorname*{argmin}_y f(y) + \frac{1}{2\gamma}\|y - x^n\|_2^2, \\ z^{n+1} &\in \operatorname*{argmin}_z g(z) + \frac{1}{2\gamma}\|2y^{n+1} - x^n - z\|_2^2, \\ x^{n+1} &= x^n + (z^{n+1} - y^{n+1}). \end{aligned}$$

The global convergence of the whole sequence $((y^n, z^n, x^n))_{n\in\mathbb{N}}$ is shown in [32, Theorem 2] for certain values of $\gamma > 0$, and is based on a descent property of the merit function

$$\mathfrak{D}_\gamma(y, z, x) := f(y) + g(z) - \frac{1}{2\gamma}\|y - z\|_2^2 + \langle x - y, z - y\rangle .$$

During the proof, which they tailored to their method, the abstract conditions in Assumption H are verified. Condition (H1) is verified in [32, equation (23)] with some constant $a > 0$ for the function $\mathfrak{D}_\gamma$ using $d^n := \|y^{n+1} - y^n\|_2$,

$$\mathfrak{D}_\gamma(y^{n+1}, z^{n+1}, x^{n+1}) + a\|y^{n+1} - y^n\|_2^2 \le \mathfrak{D}_\gamma(y^n, z^n, x^n).$$

Condition (H2) is established in [32, equation (28)] for some $b > 0$,

$$\operatorname{dist}(0, \partial\mathfrak{D}_\gamma(y^n, z^n, x^n)) \le b\|y^{n+1} - y^n\|_2,$$

using $I := \{0\}$, $\theta_0 = 1$, $b_n \equiv 1$, $\varepsilon_n \equiv 0$, and (H3) is proved by assuming the existence of a cluster point and the $\mathfrak{D}_\gamma$-attentive convergence from [32, equations (25)–(27)]. The distance condition (H4) is asserted by [32, equations (22) and (10)] and the relation in the x-update step. Condition (H5) is obviously satisfied, since we are in a setting with constant parameters. Therefore, we can apply our Theorem 10 to prove the same convergence results as in [32, Theorem 2]: $(y^n)_{n\in\mathbb{N}}$ converges and, using the same equations that realize the distance condition, convergence of $(z^n)_{n\in\mathbb{N}}$ and $(x^n)_{n\in\mathbb{N}}$ can be concluded.

Relation to [33]. In a similar way to [32], the proximal ADMM proposed in [33] can be cast in our framework. The goal is to solve the problem

$$\min_{x\in\mathbb{R}^N} h(x) + P(\mathcal{M}x),$$

with a linear mapping $\mathcal{M}$, a proper lower semicontinuous function P, and a twice continuously differentiable function h with bounded Hessian. The sufficient decrease condition is proved for the Lagrange function

$$L_\beta(x,y,z) = h(x) + P(y) - \langle z, \mathcal{M}x - y\rangle + \frac{\beta}{2}\|\mathcal{M}x - y\|_2^2$$

in [33, equation (36)] with $d^n := \|x^{n+1} - x^n\|_2$, and some $a > 0$. Different from the analysis in [32], where the relative error condition is explicit, it is implicit in [33]. Condition (H2) is verified in [33, equation (35)] for some $b > 0$, $b_n \equiv 1$, $\varepsilon_n \equiv 0$, $I = \{1\}$, and $\theta_1 = 1$. Condition (H3) is proved in [33, Theorem 2(i)]. The distance condition (H4) follows directly from [33, equations (14) and (15)], and (H5) is again obviously satisfied.

Relation to [36]. A very general multistep forward-backward scheme is proposed to solve problems in the setting of (36). The main update step is a forward-backward step, executed at an extrapolated point with gradient direction evaluated at another extrapolated point. Both of these extrapolations allow for a linear combination (possibly different ones) of finitely many preceding step directions. Global convergence and a finite length property are proved in [36, Theorem 2.2] explicitly for this algorithm for the sequence $(x^n)_{n\in\mathbb{N}}$ and $(z^n)_{n\in\mathbb{N}}$ with $z^n = (x^n, x^{n-1}, \ldots, x^{n-s+1})$ for some $s \in \mathbb{N}$. The statements that establish the conditions in Assumption H are collected in [36, (R.1)–(R.3)] in the supplementary material. The proof idea follows the concepts of the proof of iPiano [44]. The arising Lyapunov function and the product space is naturally generalized to the number of terms used in the linear combinations of the extrapolations.

REFERENCES

[1] P. A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2005), pp. 531–547.

[2] L. Ambrosio and V. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via γ-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

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

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457, https://doi.org/10.1287/moor.1100.0449.

[5] H. Attouch, J. Bolte, and B. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods*, Math. Program., 137 (2013), pp. 91–129, https://doi.org/10.1007/s10107-011-0484-9.

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

[7] J. Bolte, P. L. Combettes, and J.-C. Pesquet, *Alternating proximal algorithm for blind image recovery*, in Proceedings of the 2010 IEEE International Conference on Image Processing, IEEE Press, Piscataway, NJ, 2010, pp. 1673–1676.

[8] J. Bolte, A. Daniilidis, and A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2007), pp. 1205–1223, https://doi.org/10.1137/050644641.

[9] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572, https://doi.org/10.1137/060670080.

[10] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (2010), pp. 3319–3363.

[11] J. Bolte and E. Pauwels, *Majorization–minimization procedures and convergence of SQP methods for semi-algebraic and tame programs*, Math. Oper. Res., 41 (2016), https://doi.org/10.1287/moor.2015.0735.

[12] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.

[13] S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi, *On the Convergence of a Linesearch Based Proximal-Gradient Method for Nonconvex Optimization*, preprint, https://arxiv.org/abs/1605.03791, 2016.

[14] R. I. Boţ and E. R. Csetnek, *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, J. Optim. Theory Appl., 171 (2015), pp. 600–616.

[15] R. I. Boţ, E. R. Csetnek, and S. László, *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions*, EURO J. Comput. Optim., 4 (2016), pp. 3–25.

[16] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[17] R. Chill and M. Jendoubi, *Convergence to steady states in asymptotically autonomous semilinear evolution equations*, Nonlinear Anal., 53 (2003), pp. 1017–1039.

[18] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function*, J. Optim. Theory Appl., 162 (2013), pp. 107–132, https://doi.org/10.1007/s10957-013-0465-7.

[19] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *A block coordinate variable metric forward-backward algorithm*, J. Global Optim., 66 (2016), pp. 457–485.

[20] L. P. D. van den Dries, *Tame Topology and O-Minimal Structures*, London Math. Soc. Lecture Notes Ser. 248, Cambridge University Press, Cambridge, 1998.

[21] P. Frankel, G. Garrigos, and J. Peypouquet, *Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates*, J. Optim. Theory Appl., 165 (2014), pp. 874–900.

[22] I. Galic, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.-P. Seidel, *Towards PDE-based image compression*, in Variational, Geometric, and Level Set Methods in Computer Vision: Third International Workshop, VLSM 2005, Lecture Notes Comput. Sci. 3752, N. Paragios, O. D. Faugeras, T. Chan, and C. Schnörr, eds., Springer, Berlin, 2005, pp. 37–48.

[23] L. Grippof and M. Sciandrone, *Globally convergent block-coordinate techniques for unconstrained optimization*, Optim. Methods Softw., 10 (1999), pp. 587–637.

[24] A. Haraux and M. Jendoubi, *Convergence of solutions of second-order gradient-like systems with analytic nonlinearities*, J. Differ. Equ., 144 (1998), pp. 313–320.

[25] S. Hosseini, *Convergence of Nonsmooth Descent Methods via Kurdyka–Lojasiewicz Inequality on Riemannian Manifolds*, INS Preprint 1523, Institut für Numerische Simulation, Rheinische Friedrich–Wilhelms–Universität Bonn, Bonn, 2015.

[26] S.-Z. Huang and P. Takáč, *Convergence in gradient-like systems which are asymptotically autonomous and analytic*, Nonlinear Anal., 46 (2001), pp. 675–698.

[27] D. R. Hunter and K. Lange, *A tutorial on MM algorithms*, J. Amer. Statist. Assoc., 58 (2004), pp. 30–37.

[28] A. N. Iusem, T. Pennanen, and B. F. Svaiter, *Inexact variants of the proximal point algorithm without monotonicity*, SIAM J. Optim., 13 (2003), pp. 1080–1097.

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

[30] K. Kurdyka and A. Parusiński, *w_f-stratification of subanalytic functions and the Łojasiewicz inequality*, C. R. Acad. Paris, 318 (1994), pp. 129–133.

[31] C. Lageman, *Pointwise convergence of gradient-like systems*, Math. Nachr., 280 (2007), pp. 1543–1558.

[32] G. Li and T. Pong, *Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, Math. Program., 159 (2016), pp. 371–401.
[33] G. Li and T. K. Pong, *Global convergence of splitting methods for nonconvex composite optimization*, SIAM J. Optim., 25 (2015), pp. 2434–2460.
[34] G. Li, T. Liu, and T. K. Pong, *Peaceman–Rachford Splitting for a Class of Nonconvex Optimization Problems*, preprint, https://arxiv.org/abs/1507.00887, 2015.
[35] H. Li and Z. Lin, *Accelerated proximal gradient method for nonconvex programming*, in Adv. Neural Inf. Process. Syst. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2015, pp. 379–387.
[36] J. Liang, J. Fadili, and G. Peyré, *A multi-step inertial forward-backward splitting method for non-convex optimization*, in Adv. Neural Inf. Process. Syst. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2016, pp. 4035–4043.
[37] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles, Éditions du Centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.
[38] S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Ann. Inst. Fourier, 43 (1993), pp. 1575–1595, http://eudml.org/doc/75048.
[39] M. Mainberger and J. Weickert, *Edge-based image compression with homogeneous diffusion*, in Computer Analysis of Images and Patterns, Lecture Notes Comput. Sci. 5702, X. Jiang and N. Petkov, eds., Springer, Berlin, 2009, pp. 476–483, https://doi.org/10.1007/978-3-642-03767-2_58.
[40] B. Merlet and M. Pierre, *Convergence to equilibrium for the backward Euler scheme and applications*, Commun. Pure Appl. Anal., 9 (2010), pp. 685–702.
[41] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Norwell, MA, 2004.
[42] D. Noll, *Convergence of non-smooth descent methods using the Kurdyka–Lojasiewicz inequality*, J. Optim. Theory Appl., 160 (2013), pp. 553–572.
[43] P. Ochs, *Long Term Motion Analysis for Object Level Grouping and Nonsmooth Optimization Methods*, Ph.D. thesis, Albert-Ludwigs-Universität, Freiburg, 2015; available online from http://lmb.informatik.uni-freiburg.de//Publications/2015/Och15.
[44] P. Ochs, Y. Chen, T. Brox, and T. Pock, *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM J. Imaging Sci., 7 (2014), pp. 1388–1419, https://doi.org/10.1137/130942954.
[45] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, *On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision*, SIAM J. Imaging Sci., 8 (2015), pp. 331–372, https://doi.org/10.1137/140971518.
[46] T. Pock and S. Sabach, *Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems*, SIAM J. Imaging Sci., 9 (2016), pp. 1756–1787.
[47] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 1–17.
[48] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, New York, 1998.
[49] L. Simon, *Asymptotics for a class of non-linear evolution equations with applications to geometric problems*, Ann. Math., 118 (1983), pp. 525–571.
[50] M. Solodov and B. Svaiter, *A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.
[51] M. Solodov and B. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 50–70.
[52] M. Solodov and B. Svaiter, *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optim., 22 (2001), pp. 1013–1035.
[53] A. J. Wilkie, *Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function*, J. Amer. Math. Soc., 9 (1996), pp. 1051–1094, http://www.jstor.org/stable/2152916.
[54] S. Zavriev and F. Kostyuk, *Heavy-ball method in nonconvex optimization problems*, Comput. Math. Model., 4 (1993), pp. 336–341.