J Optim Theory Appl (2018) 177:153–180
https://doi.org/10.1007/s10957-018-1272-y

Local Convergence of the Heavy-Ball Method and iPiano for Non-convex Optimization

Peter Ochs[1]

Received: 18 October 2016 / Accepted: 21 March 2018 / Published online: 27 March 2018

Abstract A local convergence result for an abstract descent method is proved. The sequence of iterates is attracted by a local (or global) minimum, stays in its neighborhood, and converges within this neighborhood. This result allows algorithms to exploit local properties of the objective function. In particular, the abstract theory in this paper applies to the inertial forward–backward splitting method: iPiano—a generalization of the Heavy-ball method. Moreover, it reveals an equivalence between iPiano and inertial averaged/alternating proximal minimization and projection methods. Key for this equivalence is the attraction to a local minimum within a neighborhood and the fact that, for a prox-regular function, the gradient of the Moreau envelope is locally Lipschitz continuous and expressible in terms of the proximal mapping. In a numerical feasibility problem, the inertial alternating projection method significantly outperforms its non-inertial variants.

Keywords Inertial forward–backward splitting · Non-convex feasibility · Prox-regularity · Gradient of Moreau envelopes · Heavy-ball method · Alternating projection · Averaged projection · iPiano

Mathematics Subject Classification 90C26 · 90C30 · 65K05 · 49J52

1 Introduction

In non-convex optimization, we often content ourselves with local properties of the objective function. Exploiting local information, such as smoothness or prox-

Communicated by Regina S. Burachik.

✉ Peter Ochs
ochs@math.uni-sb.de

[1] Mathematical Optimization Group, Saarland University, Saarbrücken, Germany

regularity around the optimum, yields a local convergence theory. Local convergence rates can be obtained or iterative optimization algorithms can be designed, which depend on properties that are available only locally around a local optimum. For revealing such results, it is crucial that the generated sequence, once entered such a neighborhood of a local optimum, stays within this neighborhood and converges to a limit point in the same neighborhood.

As an illustrative example, suppose a point close to a local minimizer can be found by a global method, for example, by exhaustive search. In a neighborhood of the local minimizer, we can switch to a more efficient local algorithm. The local attraction of the local minimum assures that the generated sequence of iterates stays in this neighborhood, i.e., the sequence does not escape to a different local minimum, and there is no need to switch back to the (slow) global method, exhaustive search.

An important example of local properties, which we are going to exploit in this paper, is the fact that the Moreau envelope of a prox-regular function is locally well defined and its gradient is Lipschitz continuous and expressible using the proximal mapping—a result that is well known for convex functions. Locally, this result can be applied to gradient-based iterative methods for minimizing objective functions that involve a Moreau envelope of a function. We pursue this idea for the Heavy-ball method [1,2] and iPiano [3,4] (inertial version of forward–backward splitting) and obtain new algorithms for non-convex optimization such as inertial alternating/averaged proximal minimization or projection methods. The convergence result of the Heavy-ball method and iPiano translates directly to these new methods in the non-convex setting. The fact that a wide class of functions is prox-regular extends the applicability of these inertial methods significantly.

Prox-regularity was introduced in [5] and comprises primal-lower-nice (introduced by Poliquin [6]), second-order subsmooth, strongly amenable (see for instance [7]), and proper lower semi-continuous convex functions. It is known that prox-regular functions (locally) share some favorable properties of convex functions, e.g., the formula for the gradient of a Moreau envelope. Indeed, a function is prox-regular if and only if there exists a (function value attentive) localization of the subgradient mapping that is monotone up to a multiple of the identity mapping [5]. In [8], prox-regularity is key to prove local convergence of the averaged projection method using the gradient descent method, which is a result that has motivated this paper.

The convergence proof of the gradient method in [8] follows a general paradigm that is currently actively used for the convergence theory in non-convex optimization. The key is the so-called Kurdyka–Łojasiewicz (KL) property [9–13], which is known to be satisfied by semi-algebraic [14], globally subanalytic functions [15], or, more generally, functions that are definable in an o-minimal structure [13,16]. Global convergence of the full sequence generated by an abstract algorithm to a stationary point is proved for functions with the KL property. The algorithm is abstract in the sense that the generated sequence is assumed to satisfy a *sufficient descent condition*, a *relative error condition*, and a *continuity condition*; however, no generation process is specified.

The following works have also shown global convergence using the KL property or earlier versions thereof. The gradient descent method is considered in [8,17], and the proximal algorithm is analyzed in [8,18–20] and the non-smooth subgradient method in [21,22]. Convergence of forward–backward splitting (proximal gradient algorithm) is proved in [8]. Extensions to a variable metric are studied in [23] and in [24] with line search. A block coordinate descent version is considered in [25] and a block coordinate variable metric method in [26]. A flexible relative error handling of forward–backward splitting and a non-smooth version of the Levenberg–Marquardt algorithm is explored in [27]. For proximal alternating minimization, we refer to [28] for an early convergence result of the iterates and to [29] for proximal alternating linearized minimization.

Inertial variants of these algorithms have also been examined. [3] establishes convergence of an inertial forward–backward splitting algorithm, called iPiano. [3] assumes the non-smooth part of the objective to be convex, whereas [4] and [30] prove convergence in the full non-convex setting, i.e., when the algorithm is applied to minimizing the sum of a smooth non-convex function with Lipschitz gradient and a proper lower semi-continuous function. An extension to an inertial block coordinate variable metric version was studied in [31]. Bregman proximity functions are considered in [30]. A similar method was considered in [32] by the same authors. The convergence of a generic multi-step method is proved in [33] (see also [34]). A slightly weakened formulation of the popular accelerated proximal gradient algorithm from convex optimization was analyzed in [35]. Another fruitful concept from convex optimization is that of composite objective functions involving linear operators. This problem is approached in [36,37]. Key for the convergence results is usually a decrease condition on the objective function or an upper bound of the objective. The Lyapunov-type idea is studied in [37–39]. Convergence of the abstract principle of majorization minimization methods was also analyzed in a KL framework [40,41].

The global convergence theory of an unbounded memory multi-step method was proposed in [33]. Local convergence was analyzed under the additional partial smoothness assumption. In particular, local linear convergence of the iterates is established. Although the fruitful concept of partial smoothness is very interesting, in this paper, we focus on convergence results that can be inferred directly from the KL property. In the general abstract setting, local convergence rates were analyzed in [27,42] and for inertial methods in [34,42]. More specific local convergence rates can be found in [18,26,28,29,43,44].

While the abstract concept in [8] can be used to prove global convergence in the non-convex setting for the gradient descent method, forward–backward splitting, and several other algorithms, it seems to be limited to single-step methods. Therefore, [3] proved a slightly different result for abstract descent methods, which is applicable to multi-step methods, such as the Heavy-ball method and iPiano. In [31], an abstract convergence result is proved that unifies [3,4,8,27].

Contribution In this paper, we develop the *local convergence theory* for the abstract setting in [3], in analogy to the local theory in [8]. Our local convergence result shows that, for multi-step methods such as the *Heavy-ball method* or *iPiano*, a sequence that is initialized close enough to a local minimizer

- *stays in a neighborhood of the local minimum* and

- *converges to a local minimizer* instead of a stationary point.

This result allows us to apply the formula for the gradient of the Moreau envelope of a prox-regular function to all iterates, which has far-reaching consequences and has not been explored algorithmically before. We obtain *several new algorithms* for non-convex optimization problems. Conceptionally, the algorithms are known from the convex setting or from their non-inertial versions; however, *there are no guarantees for the inertial versions in the non-convex setting*.

- The Heavy-ball method applied to the sum of distance functions to prox-regular sets (resp. the sum of Moreau envelopes of prox-regular functions) coincides with the inertial averaged projection method (resp. the inertial averaged proximal minimization) for these prox-regular sets (resp. functions).
- iPiano applied to the sum of the distance function to a prox-regular set (resp. the Moreau envelope of a prox-regular function) and a simple non-convex set (resp. function) leads to the inertial alternating projection method (resp. inertial alternating proximal minimization) for these two sets (resp. functions).

Of course, these algorithms are only efficient when the associated proximal mappings or projections are simple (efficient to evaluate). Beyond these local results, we provide *global convergence guarantees* (see Proposition 5.5(2)) for the following methods:

- The (relaxed) alternating projection method for the feasibility problem of a convex set and a non-convex set.
- An inertial version of the alternating projection method (iPiano applied to the distance function to a convex set over a non-convex constraint set).
- An inertial version of alternating proximal minimization (iPiano applied to the sum of the Moreau envelope of a convex function and a non-convex function).

Moreover, we transfer *local convergence rates* depending on the KL exponent of the involved functions to the methods listed above. This result builds on a recent classification of local convergence rates depending on the KL exponent from [34,42] (which extends results from [27]).

Outline Section 2 introduces the notation and definitions that are used in this paper. In Sect. 3.1, the conditions for global convergence of abstract descent methods [3,4] are recapitulated. The main result for abstract descent methods, the *attraction of local (or global) minima*, is developed and proved in Sect. 3.2. Then, the abstract local convergence results are *verified for iPiano* (hence the Heavy-ball method) in Sect. 4. The *equivalence* to inertial averaged/alternating minimization/projection methods is analyzed in Sect. 5. Section 5.4 shows a numerical example of a *feasibility problem*.

1.1 Perspectives and Open Problems

Key for this paper is the formula for the gradient of the Moreau envelope of a function and the local capturing result that proves existence of a neighborhood of a local minimizer containing all iterates. Exemplarily, we used this formula for iPiano (a forward–backward splitting-like method). However, any scheme based on gradient steps, e.g., PALM [29], iPALM [45], variable metric iPiano [31], the quasi-Newton method in [36], applied to objective functions involving Moreau envelopes could make use of the Moreau envelope gradient formula, which will reveal connections to other algorithms. Conversely, gradient-free schemes that involve proximal steps, e.g., Peaceman–Rachford [38], Douglas–Rachford [39], ADMM [37], can be translated using the same formula. Since the Moreau envelope of a function is a special instance of an infimal convolution, the gradient formula may be generalized, leading again, to new algorithms and relations. Possibly, this requires the investigation of functions beyond prox-regularity. Finally, using the capturing result, other local properties may be explored, possibly allowing for an algorithmic realization of non-convex duality results.

2 Preliminaries

Throughout this paper, we will always work in a finite-dimensional Euclidean vector space $\mathbb{R}^N$ of dimension $N \in \mathbb{N}$, where $\mathbb{N} := \{1, 2, \ldots\}$. The vector space is equipped with the standard Euclidean norm $|\cdot|$ that is induced by the standard Euclidean inner product $|\cdot| := \sqrt{\langle\cdot,\cdot\rangle}$. The ball of radius $\varepsilon > 0$ around $\bar{x} \in \mathbb{R}^N$, we denote by $B_\varepsilon(\bar{x}) := \{x \in \mathbb{R}^N : |x - \bar{x}| \leq \varepsilon\}$.

As usual, we consider extended real-valued functions $f : \mathbb{R}^N \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, which are defined on the whole space with *domain* given by $\operatorname{dom} f := \{x \in \mathbb{R}^N : f(x) < +\infty\}$. A function is called *proper*, if $\operatorname{dom} f \neq \emptyset$. We define the *epigraph* of f as $\operatorname{epi} f := \{(x, \mu) \in \mathbb{R}^{N+1} : \mu \geq f(x)\}$. The *indicator function* δ_C of a set $C \subset \mathbb{R}^N$ is defined by $\delta_C(x) := 0$, if $x \in C$, and $\delta_C(x) := +\infty$, otherwise. A *set-valued mapping* $T : \mathbb{R}^N \rightrightarrows \mathbb{R}^M$, with $M, N \in \mathbb{N}$, is defined by its *graph* $\operatorname{Graph} T := \{(x, v) \in \mathbb{R}^N \times \mathbb{R}^M : v \in T(x)\}$. The range of a set-valued mapping is defined as $\operatorname{rge} T := \bigcup_{x \in \mathbb{R}^N} T(x)$.

A key concept in optimization and variational analysis is that of Lipschitz continuity, which is also known as strict continuity, is defined as in [7]:

Definition 2.1 (*Strict continuity* [7, Def. 9.1]) A single-valued mapping $F : D \to \mathbb{R}^M$ defined on $D \subset \mathbb{R}^N$ is *strictly continuous* at $\bar{x} \in D$, if

$$\operatorname{lip} F(\bar{x}) := \limsup_{\substack{x, x' \to \bar{x} \\ x \neq x'}} \frac{|F(x') - F(x)|}{|x' - x|}$$

is finite and $\operatorname{lip} F(\bar{x})$ is the *Lipschitz modulus* of F at $\bar{x}$. This is the same as saying F is locally Lipschitz continuous at $\bar{x}$ on D.

We introduce the term *f-attentive convergence*: A sequence $(x^k)_{k \in \mathbb{N}}$ is said to *f-converge* to $\bar{x}$, if $(x^k, f(x^k)) \to (\bar{x}, f(\bar{x}))$ as $k \to \infty$, and we write $x^k \xrightarrow{f} \bar{x}$.

Definition 2.2 (*Subdifferentials* [7, Def. 8.3]) The *Fréchet subdifferential* of f at $\bar{x} \in \operatorname{dom} f$ is the set $\widehat{\partial} f(\bar{x})$ of 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 $\widehat{\partial} f(\bar{x}) = \emptyset$. The so-called *(limiting) subdifferential* of f at $\bar{x} \in \operatorname{dom} f$ is defined by

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

and $\partial f(\bar{x}) := \emptyset$ for $\bar{x} \notin \operatorname{dom} f$.

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

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

Definition 2.3 (*Moreau envelope and proximal mapping* [7, Def. 1.22]) For $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and $\lambda > 0$, we define the *Moreau envelope*

$$e_\lambda f(x) := \inf_{w \in \mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w - x|^2\,,$$

and the *proximal mapping*

$$P_\lambda f(x) := \arg\min_{w \in \mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w - x|^2\,.$$

For a general function f, it might happen that $e_\lambda f(x)$ takes the values $-\infty$ and the proximal mapping is empty, i.e., $P_\lambda f(x) = \emptyset$. Therefore, the analysis of the Moreau envelope is usually coupled with the following property.

Definition 2.4 (*Prox-boundedness* [7, Def. 1.23]) We call $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ *prox-bounded*, if there exists $\lambda > 0$ such that $e_\lambda f(x) > -\infty$ for some $x \in \mathbb{R}^N$. The supremum of the set of all such λ is the *threshold* λ_f of prox-boundedness.

In this paper, we focus on so-called prox-regular functions. These functions have many favorable properties locally, which otherwise only convex functions exhibit.

Definition 2.5 (*Prox-regularity of functions*, [7, Def. 13.27]) A function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is *prox-regular* at $\bar{x}$ for $\bar{v}$, if f is finite and locally lsc at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$, and there exists $\varepsilon > 0$ and $\lambda > 0$ such that

$$f(x') \geq f(x) + \left\langle v, x' - x\right\rangle - \frac{1}{2\lambda}|x' - x|^2 \quad \forall x' \in B_\varepsilon(\bar{x})$$

$$\text{when } v \in \partial f(x),\ |v - \bar{v}| < \varepsilon,\ |x - \bar{x}| < \varepsilon,\ f(x) < f(\bar{x}) + \varepsilon\,.$$

When this holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be prox-regular at $\bar{x}$.

The largest value $\lambda > 0$ for which this property holds is called the *modulus of prox-regularity at* $\bar{x}$.

Definition 2.6 (*Prox-regularity of sets*, [7, Ex. 13.31]) A set C is *prox-regular* at $\bar{x}$ for $\bar{v}$ when the indicator function δ_C of the set C is prox-regular at $\bar{x}$ for $\bar{v}$. It is called *prox-regular* at $\bar{x}$, when this is true for all $\bar{v} \in \partial\delta_C(\bar{x})$.

We provide several examples to show that most functions in practice are prox-regular.

Example 2.1 A function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ is prox-regular if, for example,[1] f is

– proper lower semi-continuous (lsc) convex [7, Ex. 13.30],
– locally representable in the form $f = g - \rho|\cdot|^2$ with g being finite ($g < +\infty$) convex and $\rho > 0$ [7, Thm. 10.33],
– strongly amenable [7, Def. 10.23, Prop. 13.32] (e.g., $\mathcal{C}^2$-functions, functions of the form $g \circ F$ with F being $\mathcal{C}^2$ and g being proper lsc convex, the maximum of $\mathcal{C}^2$-function),
– lower-$\mathcal{C}^2$ [7, Def. 10.29, Prop. 13.33] (functions of the form $\max_{t\in T} f(x,t)$, where the zeroth, first, and second derivatives of $f : \mathbb{R}^N \times T \to \mathbb{R}$ w.r.t. to the first block of coordinates are continuous and T is a compact space),
– a $\mathcal{C}^2$-perturbation of a prox-regular function [7, Ex. 13.35],
– an indicator function of a closed convex set or of a strongly amenable set [7, Def. 10.23],
– the Moreau envelope of a prox-regular prox-bounded function [7, Prop. 13.37 and 13.34] (e.g., the distance function of a prox-regular set), or
– the indicator function of a closed set C and the distance function w.r.t. C is continuously differentiable on $C \setminus U$ for some open neighborhood U [46, Thm. 1.3].
– For examples of prox-regular spectral functions, we refer to [47].

Example 2.2 (Imaging problems) Several problems (image denoising, deblurring/deconvolution, zooming, depth map fusion, etc.) may be modeled as an optimization problem of the following form

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax - b|^2 + \sum_{i=1}^{N} \varphi\left(\sqrt{(Dx)_i^2 + (Dx)_{i+N}^2}\right),$$

where $A \in \mathbb{R}^{M\times N}$ (e.g., blurr operator), $b \in \mathbb{R}^M$ (e.g., blurry input image), $D \in \mathbb{R}^{2N\times N}$ (e.g., finite differences) with a continuous non-decreasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$. The objective is prox-regular, for example, under the following conditions: φ is convex and non-decreasing; $\varphi(t) = t$ (TV regularization); $\varphi(t) = \log(1+t^2)$ (student t regularization); $\varphi(t) = \log(1 + |t|)$ (at 0 where it is not $\mathcal{C}^2$, the power series of log shows that $\log(1+|t|) - |t| \in \mathcal{C}^2$; hence, φ is a $\mathcal{C}^2$-perturbation of a convex function).

[1] For the exact statements, we provide accurate references.

Example 2.3 (Support vector machine) The goal to find a linear decision function may be formulated as the following optimization problem

$$\min_{w \in \mathbb{R}^N,\, b \in \mathbb{R}} \sum_{i=1}^{M} \mathscr{L}(\langle w, z_i \rangle + b, y_i) + \varphi(w)\,,$$

where, for $i = 1, \ldots, M$, $(z_i, y_i) \in \mathbb{R}^N \times \{\pm 1\}$ is the training set, $\mathscr{L}$ is a loss, and φ is a regularizer. Examples are the hinge loss $\mathscr{L}(\bar{y}_i, y_i) = \max(0, 1 - \bar{y}_i y_i)$ (which is a maximum of $\mathcal{C}^2$-functions), the squared hinge loss, the logistic loss $\mathscr{L}(\bar{y}_i, y_i) = \log(1+e^{-\bar{y}_i y_i})$ (which are $\mathcal{C}^2$ function), etc. Prox-regular regularization functions φ are, for example, the squared ℓ_2-norm $|x|^2$, or more in general p-norms $\|x\|_p^p = \sum_{i=1}^N |x_i|^p$ with $p > 0$ ($x_i \mapsto |x_i|^p$ is $\mathcal{C}^2$ on $\mathbb{R} \setminus \{0\}$ and obviously prox-regular at $\bar{x} = 0$).

For the proof of the Lipschitz property of the Moreau envelope, it will be helpful to consider a so-called localization. A *localization* of ∂f around $(\bar{x}, \bar{v})$ is a mapping $T : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ whose graph is obtained by intersecting Graph ∂f with some neighborhood of $(\bar{x}, \bar{v})$, i.e., Graph $T =$ Graph $\partial f \cap U$ for a neighborhood U of $(\bar{x}, \bar{v})$. We talk about an *f-attentive localization* when Graph $T = \{(x, v) \in \text{Graph}\, \partial f \;:\; (x, v) \in U \text{ and } f(x) \in V\}$ for a neighborhood U of $(\bar{x}, \bar{v})$ and a neighborhood V of $f(\bar{x})$.

Finally, the convergence result we build on is only valid for functions that have the KL property at a certain point. This property is shared, for example, by semi-algebraic functions, globally subanalytic functions, or, more generally, functions definable in an o-minimal structure. For details, we refer to [12,13].

Definition 2.7 (*Kurdyka–Łojasiewicz property/KL property* [8]) Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ 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 $\phi\colon [0, \eta[\to \mathbb{R}_+$ such that

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

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

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

holds; then, the function has the Kurdyka–Łojasiewicz property at $\bar{x}$, where $\|\partial f(x)\|_- := \inf_{v \in \partial f(x)} |v|$ is the *non-smooth slope* (note: $\inf \emptyset := +\infty$).

If, additionally, the function is lsc and the property holds for each point in dom ∂f, then f is called Kurdyka–Łojasiewicz function.

If f is closed and semi-algebraic, it is well known [9,13] that f has the KL property at any point in dom ∂f, and the *desingularization function* ϕ in Definition 2.7 has the form $\phi(s) = \frac{c}{\theta} s^\theta$ for $\theta \in]0, 1]$ and some constant $c > 0$. The parameter θ is known as the *KL exponent*.

3 Abstract Convergence Result for KL Functions

In this section, we establish a local convergence result for abstract descent methods, i.e., the method is characterized by properties (H1), (H2), (H3) (see below) instead of a specific update rule. The local convergence result is inspired by a global convergence result proved in [3] for KL functions (see Theorem 3.1), which itself is motivated by a slightly different result in [8]. The abstract setting in [8] can be used to prove global and local convergence of gradient descent, proximal gradient descent, and other (single-step) methods. However, it does not apply directly to inertial variants of these methods. Therefore, in this section, we prove the required adaptation of the framework in [8] to the one in [3]. We obtain a local convergence theory that also applies to the Heavy-ball method and iPiano (see Sect. 4).

3.1 Global Convergence Results

The convergence result in [3] is based on the following three abstract conditions for a sequence $(z^k)_{k\in\mathbb{N}} := (x^k, x^{k-1})_{k\in\mathbb{N}}$ in $\mathbb{R}^{2N}$, $x^k \in \mathbb{R}^N$, $x^{-1} \in \mathbb{R}^N$. Fix two positive constants $a > 0$ and $b > 0$ and consider a proper lower semi-continuous (lsc) function $\mathcal{F}\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$. Then, the conditions for $(z^k)_{k\in\mathbb{N}}$ are as follows:

(H1) For each $k \in \mathbb{N}$, it holds that

$$\mathcal{F}(z^{k+1}) + a|x^k - x^{k-1}|^2 \le \mathcal{F}(z^k)\,.$$

(H2) For each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial\mathcal{F}(z^{k+1})$ such that

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

(H3) There exists a subsequence $(z^{k_j})_{j\in\mathbb{N}}$ such that

$$z^{k_j} \to \tilde{z} \quad \text{and} \quad \mathcal{F}(z^{k_j}) \to \mathcal{F}(\tilde{z})\,, \qquad \text{as } j \to \infty\,.$$

Theorem 3.1 (Abstract global convergence, [3, Thm. 3.7]) *Let the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, x^{k-1})_{k\in\mathbb{N}}$ satisfy (H1), (H2), and (H3) for a proper lsc function $\mathcal{F}\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$ which has the KL property at the cluster point $\tilde{z}$ specified in (H3).*

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ has finite length, i.e.,

$$\sum_{k=1}^{\infty} |x^k - x^{k-1}| < +\infty\,, \tag{2}$$

and converges to $\bar{z} = \tilde{z}$ where $\bar{z} = (\bar{x}, \bar{x})$ is a critical point of $\mathcal{F}$.

Remark 3.1 In view of the proof of this statement, obviously, the same result can be established when (H1) is replaced by $\mathcal{F}(z^{k+1}) + a|x^{k+1} - x^k|^2 \le \mathcal{F}(z^k)\,.$

3.2 Local Convergence Results

The upcoming local convergence result shows that, once entered a region of attraction (around a local minimizer), all iterates of a sequence $(z^k)_{k\in\mathbb{N}}$ satisfying (H1), (H2), and the following growth condition (H4) stay in a neighborhood of this minimum and converge to a minimizer in the same neighborhood (not just a stationary point). For the convergence to a global minimizer, the growth condition (H4) is not required.

In the following, for $z \in \mathbb{R}^{2N}$ we denote by $z_1, z_2 \in \mathbb{R}^N$ the first and second block of coordinates, i.e., $z = (z_1, z_2)$. The same holds for other vectors in $\mathbb{R}^{2N}$.

(H4) Fix $z^* \in \mathbb{R}^N$. For any $\delta > 0$, there exist $0 < \rho < \delta$ and $\nu > 0$ such that

$$z \in B_\rho(z^*)\,,\ \mathcal{F}(z) < \mathcal{F}(z^*) + \nu\,,\ y_2 \notin B_\delta(z_2^*) \ \Rightarrow\ \mathcal{F}(z) < \mathcal{F}(y) + \frac{a}{4}|z_2 - y_2|^2$$

where a is the same as in (H1)–(H3).

A simple condition that implies (H4) is provided by the following lemma:

Lemma 3.1 *Let $\mathcal{F}\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$ be proper lsc and $z^* = (x^*, x^*) \in \operatorname{dom} \mathcal{F}$ a local minimizer of $\mathcal{F}$. Suppose, for any $\delta > 0$, $\mathcal{F}$ satisfies the growth condition*

$$\mathcal{F}(y) \ge \mathcal{F}(z^*) - \frac{a}{16}|y_2 - z_2^*|^2 \quad \forall y \in \mathbb{R}^{2N},\, y_2 \notin B_\delta(z_2^*)\,.$$

Then, $\mathcal{F}$ satisfies (H4).

Proof Let $\delta > \rho$ and ν be positive numbers. For $y = (y_1, y_2) \in \mathbb{R}^{2N}$ with $y_2 \notin B_\delta(z_2^*)$ and $z = (z_1, z_2) \in B_\rho(z^*)$ such that $\mathcal{F}(z) < \mathcal{F}(z^*) + \nu$, we make the following estimation:

$$\begin{aligned}
\mathcal{F}(y) &\ge \mathcal{F}(z^*) - \frac{a}{16}|y_2 - z_2^*|^2 \\
&> \mathcal{F}(z) - \nu - \frac{a}{8}|y_2 - z_2^*|^2 + \frac{a}{16}|y_2 - z_2^*|^2 \\
&\ge \mathcal{F}(z) - \nu - \frac{a}{4}|y_2 - z_2|^2 - \frac{a}{4}|z_2 - z_2^*|^2 + \frac{a}{16}|y_2 - z_2^*|^2 \\
&\ge \mathcal{F}(z) - \frac{a}{4}|y_2 - z_2|^2 + \left(-\nu - \frac{a}{4}\rho^2 + \frac{a}{16}\delta^2\right).
\end{aligned}$$

For sufficiently small ν and ρ the term in the parenthesis becomes positive, which implies (H4). □

We need another preparatory lemma, which is proved in [3]

Lemma 3.2 ([3, Lem. 3.5]) *Let $\mathcal{F}\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$ be a proper lsc function which satisfies the Kurdyka–Łojasiewicz property at some point $z^* = (z_1^*, z_2^*) \in \mathbb{R}^{2N}$. Denote by U, η and $\phi\colon [0, \eta[\to \mathbb{R}_+$ the objects appearing in Definition 2.7 of the KL property at z^*. Let $\sigma, \rho > 0$ be such that $B_\sigma(z^*) \subset U$ with $\rho \in]0, \sigma[$.*

Furthermore, let $(z^k)_{k\in\mathbb{N}} = (x^k, x^{k-1})_{k\in\mathbb{N}}$ satisfy (H1), (H2), and

$$\forall k \in \mathbb{N}:\quad z^k \in B_\rho(z^*) \Rightarrow z^{k+1} \in B_\sigma(z^*) \text{ with } \mathcal{F}(z^{k+1}), \mathcal{F}(z^{k+2}) \ge \mathcal{F}(z^*)\,, \quad (3)$$

moreover, $z^0 = (x^0, x^{-1})$ *be such that* $\mathcal{F}(z^*) \le \mathcal{F}(z^0) < \mathcal{F}(z^*) + \eta$ *and*

$$|x^* - x^0| + \sqrt{\frac{\mathcal{F}(z^0) - \mathcal{F}(z^*)}{a}} + \frac{b}{a}\phi(\mathcal{F}(z^0) - \mathcal{F}(z^*)) < \frac{\rho}{2}\,. \tag{4}$$

Then, the sequence $(z^k)_{k\in\mathbb{N}}$ *satisfies*

$$\forall k \in \mathbb{N} : z^k \in B_\rho(z^*), \quad \sum_{k=0}^{\infty} |x^k - x^{k-1}| < \infty, \quad \mathcal{F}(z^k) \to \mathcal{F}(z^*), \text{ as } k \to \infty\,, \tag{5}$$

$(z^k)_{k\in\mathbb{N}}$ *converges to a point* $\bar{z} = (\bar{x}, \bar{x}) \in B_\sigma(z^*)$ *such that* $\mathcal{F}(\bar{z}) \le \mathcal{F}(z^*)$. *If, additionally, (H3) is satisfied, then* $0 \in \partial\mathcal{F}(\bar{z})$ *and* $\mathcal{F}(\bar{z}) = \mathcal{F}(z^*)$.

Under Assumption (H4), the following theorem establishes the local convergence result. Note that, thanks to Lemma 3.1, a global minimizer automatically satisfies (H4).

Theorem 3.2 (Abstract local convergence) *Let* $\mathcal{F} : \mathbb{R}^{2N} \to \overline{\mathbb{R}}$ *be a proper lsc function which has the KL property at some local (or global) minimizer* $z^* = (x^*, x^*)$ *of* $\mathcal{F}$. *Assume (H4) holds at* z^*.

Then, for any $r > 0$*, there exist* $u \in]0, r[$ *and* $\mu > 0$ *such that the conditions*

$$z^0 \in B_u(z^*)\,, \qquad \mathcal{F}(z^*) < \mathcal{F}(z^0) < \mathcal{F}(z^*) + \mu\,, \tag{6}$$

imply that any sequence $(z^k)_{k\in\mathbb{N}}$ *that starts at* z^0 *and satisfies (H1) and (H2) has the finite length property (2) and remains in* $B_r(z^*)$ *and converges to some* $\bar{z} \in B_r(z^*)$*, a critical point of* $\mathcal{F}$ *with* $\mathcal{F}(\bar{z}) = \mathcal{F}(z^*)$. *For r sufficiently small,* $\bar{z}$ *is a local minimizer of* $\mathcal{F}$.

Proof Let $r > 0$. Since $\mathcal{F}$ satisfied the KL property at z^*, there exist $\eta_0 \in]0, +\infty]$, $\delta \in]0, r/\sqrt{2}[$ and a continuous concave function $\phi : [0, \eta_0[\to \mathbb{R}$ such that $\phi(0) = 0$, ϕ is continuously differentiable and strictly increasing on $]0, \eta_0[$, and for all

$$z \in B_{\sqrt{2}\delta}(z^*) \cap [\mathcal{F}(z^*) < \mathcal{F}(z) < \mathcal{F}(z^*) + \eta_0]$$

the KL inequality holds. As z^* is a local minimizer, by choosing a smaller δ if necessary, one can assume that

$$\mathcal{F}(z) \ge \mathcal{F}(z^*) \quad \text{for all} \quad z \in B_{\sqrt{2}\delta}(z^*)\,. \tag{7}$$

Let $0 < \rho < \delta$ and $\nu > 0$ be the parameters appearing in (H4) with δ as in (7). We want to verify the implication in (3) with $\sigma = \sqrt{2}\delta$. Let $\eta := \min(\eta_0, \nu)$ and $k \in \mathbb{N}$. Assume $z^0, \ldots, z^k \in B_\rho(z^*)$, with $z^k =: (z_1^k, z_2^k) = (x^k, x^{k-1}) \in \mathbb{R}^{N\times 2}$ and w.l.o.g. $\mathcal{F}(z^*) < \mathcal{F}(z^0), \ldots, \mathcal{F}(z^k) < \mathcal{F}(z^*) + \eta$ (note that if $\mathcal{F}(z^k) = \mathcal{F}(z^*)$, the sequence is stationary ($x^k = x^{k+1} = x^{k+2} = \ldots$) by (H1) and the result follows directly).

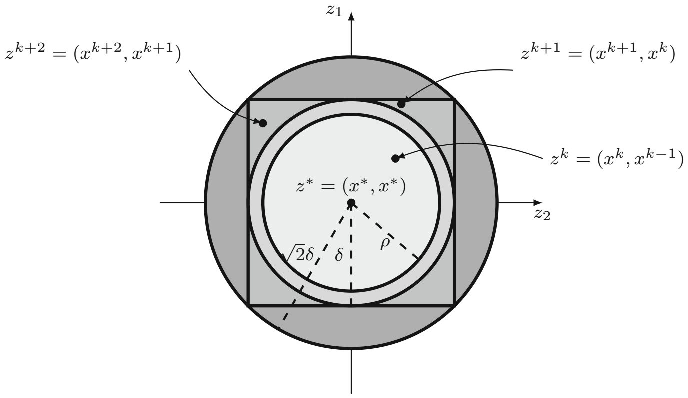

Fig. 1 An essential step of the proof of Theorem 3.2 is to show: $z^k \in B_\rho(z^*) = B_\rho(x^*, x^*)$ implies $x^{k+2}, x^{k+1} \in B_\delta(z_2^*) = B_\delta(x^*)$ which restricts z^{k+1} and z^{k+2} to the rectangle in the plot and thus to $B_{\sqrt{2}\delta}(z^*)$

See Fig. 1 for the idea of the following steps. First, note that $x^k \in B_\delta(z_2^*)$ as $z^k \in B_\delta(z^*)$. Suppose $z_2^{k+2} = x^{k+1} \notin B_\delta(z_2^*)$. Then by (H4) and (H1), we observe (use $(u+v)^2 \leq 2(u^2+v^2)$)

$$\begin{aligned}\mathcal{F}(z^k) &< \mathcal{F}(z^{k+2}) + \frac{a}{4}|x^{k-1} - x^{k+1}|^2 \\ &\leq \mathcal{F}(z^k) - a\left(|x^{k+1} - x^k|^2 + |x^k - x^{k-1}|^2\right) + \frac{a}{4}|x^{k-1} - x^{k+1}|^2 \leq \mathcal{F}(z^k)\,,\end{aligned}$$

which is a contradiction and therefore $z_2^{k+2} \in B_\delta(z_2^*)$.

Hence, we have $z^{k+1} = (x^{k+1}, x^k) \in B_{\sqrt{2}\delta}(z^*)$, due to the equivalence of norms in finite dimensions. Thanks to (7), we have $\mathcal{F}(z^{k+1}) \geq \mathcal{F}(z^*)$. In order to verify (3), we also need $\mathcal{F}(z^{k+2}) \geq \mathcal{F}(z^*)$, which can be shown analogously; however, we need to consider three iteration steps (that is the reason for the factor $\frac{a}{4}$ instead of $\frac{a}{2}$ on the right-hand side of (H4)). The assumption that $z_2^{k+3} = x^{k+2} \notin B_\delta(z_2^*)$ yields the following contradiction:

$$\begin{aligned}\mathcal{F}(z^k) &< \mathcal{F}(z^{k+3}) + \frac{a}{4}|x^{k-1} - x^{k+2}|^2 \\ &\leq \mathcal{F}(z^k) - a\left(|x^{k+2} - x^{k+1}|^2 + |x^{k+1} - x^k|^2 + |x^k - x^{k-1}|^2\right) \\ &\quad + \frac{a}{4}|x^{k-1} - x^{k+2}|^2 \\ &\leq \mathcal{F}(z^k) - a\left(|x^{k+2} - x^{k+1}|^2 + |x^{k+1} - x^k|^2 + |x^k - x^{k-1}|^2\right) \\ &\quad + \frac{a}{4}\left(2|x^{k+2} - x^{k+1}|^2 + 4|x^{k+1} - x^k|^2 + 4|x^k - x^{k-1}|^2\right) \leq \mathcal{F}(z^k)\,.\end{aligned}$$

Thus, $\mathcal{F}(z^{k+1}), \mathcal{F}(z^{k+2}) \geq \mathcal{F}(z^*)$ holds, which is property (3) with $\sigma = \sqrt{2}\delta$.

Now, choose $u, \mu > 0$ in (6) such that $\mu < \eta\,,\ u < \frac{\rho}{6}\,,\ \sqrt{\frac{\mu}{a}} + \frac{b}{a}\phi(\mu) < \frac{\rho}{3}$. If z^0 satisfies (6), we have

$$|x^* - x^0| + \sqrt{\frac{\mathcal{F}(z^0) - \mathcal{F}(z^*)}{a}} + \frac{b}{a}\phi(\mathcal{F}(z^0) - \mathcal{F}(z^*)) < \frac{\rho}{2},$$

which is (4) with μ in place of η. Using Lemma 3.2, we conclude that the sequence has finite length, remains in $B_\rho(z^*)$, converges to $\bar{z} \in B_\sigma(z^*)$, and $\mathcal{F}(z^k) \to \mathcal{F}(z^*)$ and $\mathcal{F}(\bar{z}) \le \mathcal{F}(z^*)$, which is only allowed for $\mathcal{F}(\bar{z}) = \mathcal{F}(z^*)$. Therefore, the sequence also has property (H3), and thus, $\bar{z}$ is a critical point of $\mathcal{F}$. Property (7) shows that $\bar{z}$ is a local minimizer for sufficiently small r. □

Remark 3.2 The assumption in (H4) and Lemma 3.1 only restrict the behavior of the function along the second block of coordinates of $z = (z_1, z_2) \in \mathbb{R}^{2N}$. This makes sense, because, for sequences that we consider, the first and second block depend on each other.

Remark 3.3 Unlike Theorem 3.1, the local convergence theorem (Theorem 3.2) does not require (H3) explicitly. If Theorem 3.1 assumes the KL property at some z^* (not the cluster point $\tilde{z}$ of (H3)), convergence to a point $\bar{z}$ in a neighborhood of z^* with $\mathcal{F}(\bar{z}) \le \mathcal{F}(z^*)$ can be shown. However, $\mathcal{F}(\bar{z}) < \mathcal{F}(z^*)$ might happen, which disproves $\mathcal{F}$-attentive convergence of $z^k \to \bar{z}$; thus, $\bar{z}$ would not be a critical point. Assuming $\tilde{z} = z^*$ by (H3) assures the $\mathcal{F}$-attentive convergence, and thus, $\bar{z}$ is a critical point. Because of the local minimality of z^* in Theorem 3.2, $\mathcal{F}(\bar{z}) < \mathcal{F}(z^*)$ cannot occur, and therefore, (H3) is implied.

Before deriving the convergence rates, we apply Theorem 3.2 and Lemma 3.1 to show a useful example of a feasibility problem.

Example 3.1 (Semi-algebraic feasibility problem) Let $S_1, \ldots, S_M \subset \mathbb{R}^N$ be semi-algebraic sets such that $\bigcap_{i=1}^M S_i \neq \emptyset$ and let $F\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be given by $F(x) = \frac{1}{2}\sum_{i=1}^M \operatorname{dist}(x, S_i)^2$. For a constant $c \ge 0$, we consider the function $\mathcal{F}(z) = \mathcal{F}(z_1, z_2) = F(z_1) + c|z_1 - z_2|^2$. Suppose $z^* = (x^*, x^*)$ is a global minimizer of $\mathcal{F}$, i.e., $x^* \in \bigcap_{i=1}^M S_i$. Then, for $z^0 = (x^0, x^{-1})$ sufficiently close to z^*, any algorithm that satisfies (H1) and (H2) and starts at z^0 generates a sequence that remains in a neighborhood of z^* has the finite length property and converges to a point $\bar{z} = (\bar{x}, \bar{x})$ with $\bar{x} \in \bigcap_{i=1}^M S_i$.

Finally, we complement our local convergence result by the convergence rate estimates from [34,42]. Assuming the objective function is semi-algebraic, in [34, Thm.s 2 and 4] which build on [27, Thm. 3.4], a list of qualitative convergence rate estimates in terms of the KL exponent is proved. For estimations on the KL exponent, the interested reader is referred to [42,48–50], which include estimations of the KL exponent for convex polynomials, functions that can be expressed as the maximum of finitely many polynomials, functions that can be expressed as supremum of a collection of polynomials over a semi-algebraic compact set under suitable regularity assumptions, and relations to the Luo–Tseng error bound.

Theorem 3.3 (Convergence rates) *Let $(z^k)_{k\in\mathbb{N}} = (x^k, x^{k-1})_{k\in\mathbb{N}}$ satisfy (H1), (H2), and (H3) for a proper lsc function $\mathcal{F}\colon \mathbb{R}^{2N} \to \overline{\mathbb{R}}$ with KL exponent θ, which has the KL property at the critical point $\tilde{z} = z^*$ specified in (H3).*

1. *If* $\theta = 1$, *then* z^k *converges to* z^* *in a finite number of iterations.*
2. *If* $\frac{1}{2} \leq \theta < 1$, *then* $\mathcal{F}(z^k) \to \mathcal{F}(z^*)$ *and* $x^k \to x^*$ *linearly.*
3. *If* $0 < \theta < \frac{1}{2}$, *then* $\mathcal{F}(z^k) - \mathcal{F}(z^*) \in O(k^{\frac{1}{2\theta-1}})$ *and* $|x^k - x^*| \in O(k^{\frac{\theta}{2\theta-1}})$.

Proof Using Theorem 3.1, $(z^k)_{k\in\mathbb{N}}$ converges to z^* and $\mathcal{F}(z^k) \to \mathcal{F}(z^*)$ as $k \to \infty$. W.l.o.g. we can assume that $\mathcal{F}(z^k) > \mathcal{F}(z^*)$ for all $k \in \mathbb{N}$. By convergence of $(z^k)_{k\in\mathbb{N}}$ and (H1), there exists k_0 such that the KL inequality (1) with $f = \mathcal{F}$ holds for all $k \geq k_0$. Let U, ϕ, η be the objects appearing in Definition 2.7. Now, using $(u+v)^2 \leq 2(u^2+v^2)$ for $u, v \in \mathbb{R}$ to bound the terms on the right-hand side of (H2) and substituting (H1) into the resulting terms, the squared KL inequality (1) at index k yields

$$\frac{b^2}{2a}\big(\phi'(\mathcal{F}(z^k) - \mathcal{F}(z^*))\big)^2\big(\mathcal{F}(z^{k-1}) - \mathcal{F}(z^{k+1})\big) \geq 1\,.$$

As $\phi'(s) = cs^{\theta-1}$ is non-increasing for $\theta \in [0,1]$, we have $\phi'(\mathcal{F}(z^k) - \mathcal{F}(z^*)) \leq \phi'(\mathcal{F}(z^{k+1}) - \mathcal{F}(z^*))$. The remainder of the proof is identical to [34] starting from [34, Inequality (7)], which yields the rates for $(\mathcal{F}(z^k))_{k\in\mathbb{N}}$.

In the following, we prove the rates for $(x^k)_{k\in\mathbb{N}}$. We make use of an intermediate result from the proof of [3, Lem. 3.5] (cf. Lemma 3.2). The starting point is [3, Inequality (6)], restricted to terms with index $k \geq K$, $K \in \mathbb{N}$:

$$\sum_{k\geq K} |x^k - x^{k-1}| \leq \frac{1}{2}|x^K - x^{K-1}| + \frac{b}{a}\phi(\mathcal{F}(z^K) - \mathcal{F}(z^*))\,.$$

The triangle inequality shows that the left-hand side is an upper bound for $|x^K - x^*|$. Using (H1) to bound the right-hand side yields:

$$|x^K - x^*| \leq \sum_{k\geq K} |x^k - x^{k-1}| \leq c''\left(\phi(\mathcal{F}(z^K) - \mathcal{F}(z^*)) + \sqrt{\mathcal{F}(z^K) - \mathcal{F}(z^*)}\right)$$

for some constant $c'' > 0$. If $\theta \in [\frac{1}{2}, 1[$, for $\mathcal{F}(z^K) - \mathcal{F}(z^*) < 1$, the second term upper-bounds the first one, and $\mathcal{F}(z^K) \to \mathcal{F}(z^*)$ is linear. For $\theta \in]0, \frac{1}{2}[$ the first term dominates, hence $|x^K - x^*| \in O(\phi(\mathcal{F}(z^K) - \mathcal{F}(z^*)))$. □

4 Local and Global Convergence of iPiano

In this section, we briefly review the method iPiano and verify that the abstract convergence results from Sect. 3 hold for this algorithm.

iPiano applies to structured non-smooth and non-convex optimization problems with a proper lower semi-continuous (lsc) function $h\colon \mathbb{R}^N \to \overline{\mathbb{R}}$, $N \geq 1$:

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

that satisfies the following assumption.

Assumption 1 For $U \subset \mathbb{R}^N$, the following properties hold:

- The function $f : U \to \mathbb{R}$ is assumed to be C^1-smooth (possibly non-convex) with L-Lipschitz continuous gradient on dom $g \cap U$, $L > 0$.
- The function $g : U \to \overline{\mathbb{R}}$ is proper, lsc, possibly non-smooth and non-convex, simple and prox-bounded.
- The function h restricted to U is bounded from below by $\underline{h} > -\infty$ and coercive, i.e., $|x| \to \infty$ with $x \in U$ implies that $h(x) \to \infty$.

Remark 4.1 As we will use Assumption 1, either with $U = \mathbb{R}^N$ or with $U = B_{r'}(x^*)$ for some $r' > 0$, the coercivity assumption reduces either to the usual definition ($U = \mathbb{R}^N$) or is empty (since $B_{r'}(x^*)$ is bounded). The coercivity property could be replaced by the assumption that the sequence that is generated by the algorithm is bounded.

Remark 4.2 Simple refers to the fact that the associated proximal map can be solved efficiently for the global optimum.

iPiano is outlined in Algorithm 1. For $g = 0$, iPiano coincides with the Heavy-ball method (inertial gradient descent or gradient descent with momentum).

In [4], functions g that are semi-convex received special attention. The resulting step size restrictions for semi-convex functions g are similar to those of convex functions. A function is said to be semi-convex with modulus $m \in \mathbb{R}$, if m is the largest value such that $g(x) - \frac{m}{2}|x|^2$ is convex. For convex functions, $m = 0$ holds, and for strongly convex functions $m > 0$. We assume $m < L$. According to [7, Thm. 10.33], saying a function g is (locally) semi-convex on an open set $V \subset$ dom g is the same as saying g is lower-$\mathcal{C}^2$ on V. Nevertheless, the function g does not need to be semi-convex. This property is just used to improve the bounds on the step size parameters.

Remark 4.3 For simplicity, we describe the constant step size version of iPiano. However, all results in this paper are also valid for the backtracking line-search version of iPiano.

Algorithm 1 *iPiano*

- ***Optimization problem:***

$$(8)\ \textit{with Assumption } 1 \textit{ for} \begin{cases} U = \mathbb{R}^N \\ U = B_{r'}(x^*) \ \textit{ for a local minimizer } x^* \textit{ and } r' > 0. \end{cases}$$

- ***Initialization***: *Choose a starting point* $x^0 \in \operatorname{dom} h \cap U$ *and set* $x^{-1} = x^0$.
- ***Iterations*** $(k \geq 0)$*: Update:*

$$\begin{aligned} y^k &= x^k + \beta(x^k - x^{k-1}) \\ x^{k+1} &\in \arg\min_{x \in \mathbb{R}^N} \; g(x) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\alpha}|x - y^k|^2 . \end{aligned} \tag{9}$$

- ***Parameter setting***: *See Table 1.*

Table 1 Convergence of iPiano as stated in Corollaries 4.1, 4.2, and 4.3 is guaranteed for the parameter settings listed in this table (for g convex, see [3, Algorithm 2], otherwise see [4, Algorithm 3])

Method	f	g	α	β
Gradient descent	$f \in \mathcal{C}^{1+}$	$g \equiv 0$	$\alpha \in]0, \frac{2}{L}[$	$\beta = 0$
Heavy-ball method	$f \in \mathcal{C}^{1+}$	$g \equiv 0$	$\alpha \in]0, \frac{2(1-\beta)}{L}[$	$\beta \in [0, 1[$
PPA	$f \equiv 0$	g convex	$\alpha > 0$	$\beta = 0$
FBS	$f \in \mathcal{C}^{1+}$	g convex	$\alpha \in]0, \frac{2}{L}[$	$\beta = 0$
FBS (non-convex)	$f \in \mathcal{C}^{1+}$	g non-convex	$\alpha \in]0, \frac{1}{L}[$	$\beta = 0$
iPiano	$f \in \mathcal{C}^{1+}$	g convex	$\alpha \in]0, \frac{2(1-\beta)}{L}[$	$\beta \in [0, 1[$
iPiano	$f \in \mathcal{C}^{1+}$	g non-convex	$\alpha \in]0, \frac{(1-2\beta)}{L}[$	$\beta \in [0, \frac{1}{2}[$
iPiano	$f \in \mathcal{C}^{1+}$	g m-semi-convex	$\alpha \in]0, \frac{2(1-\beta)}{L-m}[$	$\beta \in [0, 1[$

Note that for local convergence, also the required properties of f and g are required to hold only locally. iPiano has several well-known special cases, such as the gradient descent method, Heavy-ball method, proximal point algorithm (PPA), and forward–backward splitting (FBS). $\mathcal{C}^{1+}$ denotes the class of functions whose gradient is strictly continuous (Lipschitz continuous)

The following convergence results hold for iPiano.

Corollary 4.1 (Global convergence of iPiano [4, Thm. 6.6]) *Let $(x^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 with $U = \mathbb{R}^N$. Then, the sequence $(z^k)_{k\in\mathbb{N}}$ with $z^k = (x^k, x^{k-1})$ satisfies (H1), (H2), (H3) for the function (for some $\kappa > 0$)*

$$H_\kappa : \mathbb{R}^{2N} \to \mathbb{R} \cup \{\infty\}, \quad (x, y) \mapsto h(x) + \kappa |x - y|^2 . \tag{10}$$

Moreover, if $H_\kappa(x, y)$ has the Kurdyka–Łojasiewicz property at a cluster point $z^* = (x^*, x^*)$, then the sequence $(x^k)_{k\in\mathbb{N}}$ has the finite length property, $x^k \to x^*$ as $k \to \infty$, and z^* is a critical point of H_κ; hence, x^* is a critical point of h.

Corollary 4.2 (Local convergence of iPiano) *Let $(x^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 with $U = B_{r'}(x^*)$ for some $r' > 0$, where x^* is a local (or global) minimizer of h. Then, $z^* = (x^*, x^*)$ is a local (or global) minimizer of H_κ (defined in (10)). Suppose (H4) holds at z^* and H_κ has the KL property at z^*.*

Then, for any $r > 0$ (in particular for $r = r'$), there exist $u \in]0, r[$ and $\mu > 0$ such that the conditions

$$x^0 \in B_u(x^*) , \qquad h(x^*) < h(x^0) < h(x^*) + \mu ,$$

imply that the sequence $(x^k)_{k\in\mathbb{N}}$ has the finite length property and remains in $B_r(x^*)$ and converges to some $\bar{x} \in B_r(x^*)$, a critical point of h that satisfies $h(\bar{x}) = h(x^*)$. For r sufficiently small, $\bar{z}$ is a local minimizer of h.

Proof Corollary 4.1 shows that Algorithm 1 generates a sequence that satisfies (H1), (H2), (H3) with H_κ. Therefore, obviously, Theorem 3.2 can be applied. □

Corollary 4.3 (Convergence rates for iPiano) *Let $(x^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 and set $z^k := (x^k, x^{k-1})$. If H_κ, defined in (10), has the KL property at $z^* = (x^*, x^*)$ specified in (H3) with KL exponent θ, then the following rates of convergence hold for some $C > 0$:*

1. *If $\theta = 1$, then x^k converges to x^* in a finite number of iterations.*
2. *If $\frac{1}{2} \leq \theta < 1$, then $h(x^k) \to h(x^*)$ and $x^k \to x^*$ linearly.*
3. *If $0 < \theta < \frac{1}{2}$, then $h(x^k) - h(x^*) \in C(k^{\frac{1}{2\theta-1}})$ and $|x^k - x^*| \in O(k^{\frac{\theta}{2\theta-1}})$.*

Proof Corollary 4.1 shows that Algorithm 1 generates a sequence that satisfies (H1), (H2), (H3) for H_κ. Therefore, the statement follows from Theorem 3.3 and the facts that $H_\kappa(x^*, x^*) = h(x^*)$ and $h(x^k) \leq H_\kappa(x^k, x^{k-1})$. □

Remark 4.4 In [42, Thm. 3.6], Li and Pong show that if h has the KL exponent $\theta \in]0, \frac{1}{2}]$ at x^*, then H_κ has the same KL exponent at $z^* = (x^*, x^*)$.

5 Inertial Averaged/Alternating Minimization

In this section, we transfer the convergence result developed for iPiano in Sect. 4 to various non-convex settings (Sects. 5.1, 5.2, 5.3). This yields inertial algorithms for non-convex problems that are known from the convex setting as averaged or alternating proximal minimization (or projection) methods. Key for the generalization to the non-convex and inertial setting is an explicit formula for the gradient of the Moreau envelope of a prox-regular function (Proposition 5.2), which is well known for convex functions (Proposition 5.1), and the local convergence results in Theorem 3.2. For completeness, we state the formula in the convex setting, before we devote ourselves to the prox-regular setting.

Proposition 5.1 ([51, Prop. 12.29]) *Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper lower semicontinuous (lsc) convex function and $\lambda > 0$. Then, $e_\lambda f$ is continuously differentiable and has the λ^{-1}-Lipschitz continuous gradient*

$$\nabla e_\lambda f(x) = \frac{1}{\lambda}(x - P_\lambda f(x))\,. \tag{11}$$

Proposition 5.2 *Suppose that $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ is prox-regular at $\bar{x}$ for $\bar{v} = 0$ and that f is prox-bounded. Then for all $\lambda > 0$ sufficiently small, there is a neighborhood of $\bar{x}$ on which*

1. *$P_\lambda f$ is monotone, single-valued and Lipschitz continuous and $P_\lambda f(\bar{x}) = \bar{x}$.*
2. *$e_\lambda f$ is differentiable with $\nabla(e_\lambda f)(\bar{x}) = 0$, in fact $\nabla(e_\lambda f)$ is strictly continuous with*
$$\nabla e_\lambda f = \lambda^{-1}(I - P_\lambda f) = (\lambda I + T^{-1})^{-1} \tag{12}$$
for an f-attentive localization T of ∂f at $(\bar{x}, 0)$, where I denotes the identity mapping. This localization can be chosen so that the set $U_\lambda := \mathrm{rge}\,(I + \lambda T)$ serves for all $\lambda > 0$ sufficiently small as a neighborhood of $\bar{x}$ on which these properties hold.

3. *There is a neighborhood of $\bar{x}$ on which for small enough λ the local Lipschitz constant of $\nabla e_\lambda f$ is λ^{-1}. If λ_0 is the modulus of prox-regularity at $\bar{x}$, then $\lambda \in]0, \lambda_0/2[$ is a sufficient condition.*
4. *Any point $\tilde{x} \in U_\lambda$ with $\nabla e_\lambda f(\tilde{x}) = 0$ is a fixed point of $P_\lambda f$ and a critical point of f.*

Proof While Items 1 and 2 are proved in [7, Prop. 13.37], Items 3 (estimation of the local Lipschitz constant) and 4 are not explicitly verified. To prove Items 3 and 4, we develop the basic objects that are required in the same way as [7, Prop. 13.37]. Thus, the first part of the proof coincides with [7, Prop. 13.37].

Without loss of generality, we can take $\bar{x} = 0$. As f is prox-bounded, the condition for prox-regularity may be taken to be global, cf. [7, Prop. 8.46(f)], i.e., there exists $\varepsilon > 0$ and $\lambda_0 > 0$ such that

$$f(x') > f(x) + \langle v, x' - x \rangle - \frac{1}{2\lambda_0}|x' - x|^2 \quad \forall x' \neq x \tag{13}$$

$$\text{when } v \in \partial f(x),\ |v| < \varepsilon,\ |x| < \varepsilon,\ f(x) < f(0) + \varepsilon\,. \tag{14}$$

Let $T : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ be the f-attentive localization of ∂f specified in (14), i.e., defined by Graph $T = \{(x, v) : v \in \partial f(x),\ |v| < \varepsilon,\ |x| < \varepsilon,\ f(x) < f(0) + \varepsilon\}$. Inequality (13) is valid for any $\lambda \in]0, \lambda_0[$. Setting $u = x + \lambda v$, inequality (13) (with λ instead of λ_0) implies $f(x') + \frac{1}{2\lambda}|x' - u|^2 > f(x) + \frac{1}{2\lambda}|x - u|^2$. Therefore, $P_\lambda f(x + \lambda v) = \{x\}$ when $v \in T(x)$. In general, for any u sufficiently close to 0, thanks to Fermat's rule on the minimization problem of $P_\lambda f(u)$, we have for any $x \in P_\lambda f(u)$ that $v = (u - x)/\lambda \in T(x)$ holds. Thus, $U_\lambda = \operatorname{rge}(I + \lambda T)$ is a neighborhood of 0 on which $P_\lambda f$ is single-valued and coincides with $(I + \lambda T)^{-1}$.

3. Let $u = x + \lambda v$ and $u' = x' + \lambda v'$ be any two elements in U_λ such that $x = P_\lambda f(u)$ and $x' = P_\lambda f(u')$. Then, (x, v) and (x', v') belong to Graph T. Thus, we can add two copies of (13) where in the second copy the roles of x and x' are swapped. This sum yields for any $\lambda_1 \in]0, \lambda_0[$ instead of λ_0 in (13):

$$0 \geq \langle v - v', x' - x \rangle - \frac{1}{\lambda_1}|x' - x|^2\,. \tag{15}$$

We substitute v with $(u - x)/\lambda$ and v' with $(u' - x')/\lambda$ which yields

$$\begin{aligned} 0 &\leq \frac{1}{\lambda_1}|x' - x|^2 + \frac{1}{\lambda}\langle (u' - x') - (u - x), x' - x \rangle \\ &= \frac{1}{\lambda}\langle u' - u, x' - x \rangle + \left(\frac{1}{\lambda_1} - \frac{1}{\lambda}\right)|x' - x|^2 \end{aligned}$$

or, equivalent to that $\langle u' - u, x' - x \rangle \geq (1 - \frac{\lambda}{\lambda_1})|x' - x|^2$.

This expression helps to estimate the local Lipschitz constant of the gradient of the Moreau envelope. Using the closed-form description of $\nabla e_\lambda f$ on U_λ, we verify the

λ^{-1}-Lipschitz continuity of $\nabla e_\lambda f$ as follows, when $\lambda \le \frac{1}{2}\lambda_1$:

$$\begin{aligned}\lambda^2 |\nabla e_\lambda f(u) - \nabla e_\lambda f(u')|^2 &= |(u-u') - (P_\lambda f(u) - P_\lambda f(u'))|^2 \\ &= |x - x'|^2 - 2\left\langle u-u', x-x'\right\rangle + |u-u'|^2 \\ &\le (2\tfrac{\lambda}{\lambda_1} - 1)|x-x'|^2 + |u-u'|^2 \le |u-u'|^2\end{aligned}$$

4. Now, let $\tilde{x} \in U_\lambda$ be a point for which $\nabla e_\lambda f(\tilde{x}) = 0$ holds. Then, according to (12), we have $\tilde{x} = P_\lambda f(\tilde{x})$ or $\tilde{x} = (I + \lambda T)^{-1}(\tilde{x})$ for the localization selected above. Inverting the mapping shows that $\tilde{x} \in \tilde{x} + \lambda T(\tilde{x})$, which implies that $0 \in T(\tilde{x})$, thus $0 \in \partial f(\tilde{x})$. □

Remark 5.1 The proof of Item (iii) of Proposition 5.2 is motivated by a similar derivation for distance functions and projection operators in [52]. See [53], for a recent analysis of the differential properties of the Moreau envelope in the infinite- dimensional setting.

5.1 Heavy-Ball Method on the Moreau Envelope

Proposition 5.3 (Inertial proximal minimization method) *Let the function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be prox-regular at x^* for $v^* = 0$ with modulus $\lambda_0 > 0$ and prox-bounded with threshold $\lambda_f > 0$. Let $0 < \lambda < \min(\lambda_f, \lambda_0/2)$, $\beta \in [0,1[$, and $\alpha \in]0, 2(1-\beta)\lambda[$. Suppose that $h = e_\lambda f$ has a local minimizer x^* and H_κ, defined in (10), satisfies (H4) and the KL property at (x^*, x^*). Let $x^0 = x^{-1}$ with $x^0 \in \mathbb{R}^N$ and $(x^k)_{k\in\mathbb{N}}$ be generated by*

$$x^{k+1} \in (1 - \alpha\lambda^{-1})x^k + \alpha\lambda^{-1} P_\lambda f(x^k) + \beta(x^k - x^{k-1})\,.$$

If x_0 is sufficiently close to x^*, then sequence $(x^k)_{k\in\mathbb{N}}$

- *is uniquely determined,*
- *has the finite length property,*
- *remains in a neighborhood of x^*,*
- *and converges to a critical point $\tilde{x}$ of f with $f(\tilde{x}) = f(x^*)$.*

If f is convex, and $\lambda > 0$, $\beta \in [0,1[$, and $\alpha \in]0, 2(1-\beta)\lambda[$, then $(x^k)_{k\in\mathbb{N}}$ has finite length and converges to a global minimizer $\tilde{x}$ of f for any $x^0 \in \mathbb{R}^N$.

Proof The statement is an application of the results for the Heavy-ball method (i.e., (8) with $g \equiv 0$) to the Moreau envelope $e_\lambda f$ of the function f. Note that H_κ inherits the KL property from h (see Remark 4.4).

Since f is prox-bounded with threshold λ_f, the function is bounded from below and coercive for $\lambda < \lambda_f$. As $\lambda < \lambda_0/2$, Proposition 5.2 can be used to conclude that there exists a neighborhood U_λ of x^* such that $e_\lambda f$ is differentiable on U_λ and $\nabla e_\lambda f$ is λ^{-1}-Lipschitz continuous.

There exists a neighborhood $U \subset U_\lambda$ of x^* which contains x_0 and Corollary 4.2 can be applied. Therefore, the Heavy-ball method (Algorithm 1 with $g \equiv 0$) with

$0 < \alpha < 2(1-\beta)\lambda$ and $\beta \in [0, 1[$ generates a sequence $(x^k)_{k\in\mathbb{N}}$ that lies in U. Using the formula in (12), the update step of the Heavy-ball method applied to $e_\lambda f$ reads as follows:

$$\begin{aligned} x^{k+1} &= x^k - \alpha \nabla e_\lambda f(x^k) + \beta(x^k - x^{k-1}) \\ &= x^k - \alpha\lambda^{-1}(x^k - P_\lambda f(x^k)) + \beta(x^k - x^{k-1}) \\ &= (1-\alpha\lambda^{-1})x^k + \alpha\lambda^{-1} P_\lambda f(x^k) + \beta(x^k - x^{k-1}) \,. \end{aligned}$$

By Proposition 5.21, $P_\lambda f$ is single-valued, and by Proposition 5.24, $0 \in \partial f(\tilde{x})$. The remaining statements follow from Corollary 4.2.

The statement about convex functions follows analogously by using Proposition 5.1 instead of Proposition 5.2 and Corollary 4.1 instead of Corollary 4.2. □

Remark 5.2 Corollary 4.3 provides convergence rates for Proposition 5.3.

Remark 5.3 The question, whether $h = e_\lambda f$ has the KL property, if f has the KL property, has been analyzed for convex functions in [42]. For non-convex functions, this is a non-trivial open problem.

5.2 Heavy-Ball Method on the Sum of Moreau Envelopes

Proposition 5.4 (Inertial averaged proximal minimization method) *Suppose $f_i : \mathbb{R}^N \to \overline{\mathbb{R}}$, $i = 1, \dots, M$ are prox-regular functions at x^* for $v^* = 0$ with modulus $\lambda_0 > 0$ and prox-bounded with threshold $\lambda_f > 0$. Let $0 < \lambda < \min(\lambda_f, \lambda_0/2)$, $\beta \in [0, 1[$, and $\alpha \in]0, 2(1-\beta)\lambda[$. Suppose that $h = \sum_{i=1}^M e_\lambda f_i$ has a local minimizer x^* and H_κ, defined in (10), satisfies (H4) and the KL property at (x^*, x^*). Let $x^0 = x^{-1}$ with $x^0 \in \mathbb{R}^N$ and $(x^k)_{k\in\mathbb{N}}$ be generated by*

$$x^{k+1} \in (1-\alpha\lambda^{-1})x^k + \frac{\alpha}{M}\lambda^{-1}\sum_{i=1}^{M} P_\lambda f_i(x^k) + \beta(x^k - x^{k-1}) \,.$$

If x_0 is sufficiently close to x^*, then sequence $(x^k)_{k\in\mathbb{N}}$

- *is uniquely determined,*
- *has the finite length property,*
- *remains in a neighborhood of x^*,*
- *and converges to a critical point $\tilde{x}$ of h with $h(\tilde{x}) = h(x^*)$.*

If all f_i are convex, and $\lambda > 0$, $\beta \in [0, 1[$, and $\alpha \in]0, 2(1-\beta)\lambda[$, then $(x^k)_{k\in\mathbb{N}}$ has finite length and converges to a global minimizer $\tilde{x}$ of h for any $x^0 \in \mathbb{R}^N$.

Proof The proof is analogously to that of Proposition 5.3 except for the fact that the Heavy-ball method is applied to $\sum_{i=1}^M e_\lambda f_i$:

$$\begin{aligned} x^{k+1} &= x^k - \frac{\alpha}{M}\sum_{i=1}^M \nabla e_\lambda f_i(x^k) + \beta(x^k - x^{k-1}) \\ &= x^k - \frac{\alpha}{M}\lambda^{-1}\sum_{i=1}^M (x^k - P_\lambda f_i(x^k)) + \beta(x^k - x^{k-1}) \\ &= (1-\alpha\lambda^{-1})x^k + \frac{\alpha}{M}\lambda^{-1}\sum_{i=1}^M P_\lambda f_i(x^k) + \beta(x^k - x^{k-1})\,. \end{aligned}$$

Instead of scaling the feasible range of step sizes for α, the scaling $\frac{1}{M}$ is included in the update formula. □

Remark 5.4 Corollary 4.3 provides convergence rates for Proposition 5.4.

Remark 5.5 In contrast to Proposition 5.3, the sequence of iterates converges to a point $\tilde{x}$ for which $\sum_{i=1}^M \nabla e_\lambda f_i(\tilde{x}) = 0$ holds. We cannot directly conclude that $0 \in \partial(\sum_i f_i)(\tilde{x})$. However, if $\nabla e_\lambda f_i(\tilde{x}) = 0$ for all $i = 1, \ldots, M$, then under suitable qualification and regularity conditions (see [7, Cor. 10.9]), we can conclude that $\tilde{x}$ is a critical point of $\sum_{i=1}^M f_i$.

Example 5.1 (Inertial averaged projection method for the semi-algebraic feasibility problem) The algorithm described in Proposition 5.4 can be used to solve the semi-algebraic feasibility problem of Example 3.1. The conditions in Example 3.1 are satisfied.

5.3 iPiano on an Objective Involving a Moreau Envelope

Proposition 5.5 (Inertial alternating proximal minimization method) *Suppose $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is prox-regular at x^* for $v^* = 0$ with modulus $\lambda_0 > 0$ and prox-bounded with threshold $\lambda_f > 0$. Let $0 < \lambda < \min(\lambda_f, \lambda_0/2)$. Moreover, suppose that $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is proper, lsc, and simple. Let $x^0 = x^{-1}$ with $x^0 \in \mathbb{R}^N$ and let the sequence $(x^k)_{k\in\mathbb{N}}$ be generated by the following update rule*

$$x^{k+1} \in P_{\alpha g}\left((1-\alpha\lambda^{-1})x^k + \alpha\lambda^{-1}P_\lambda f(x^k) + \beta(x^k - x^{k-1})\right)\,.$$

We obtain the following cases of convergence results:

1. *Assume that $h = g + e_\lambda f$ has a local minimizer x^* and H_κ, defined in (10), satisfies (H4) and the KL property at (x^*, x^*). If x_0 is sufficiently close to x^*, and α, β are selected according the property of g in one of the last three rows of Table 1 with $L = \lambda^{-1}$, then the sequence $(x^k)_{k\in\mathbb{N}}$*
   - *has the finite length property,*

- *remains in a neighborhood of* x^**,*
- *and converges to a critical point* $\tilde{x}$ *of* h *with* $h(\tilde{x}) = h(x^*)$*.*

2. *Assume that* f *is convex,* $h = g + e_\lambda f$ *and* x^* *is a cluster point of* $(x^k)_{k\in\mathbb{N}}$*. Suppose* H_κ*, defined in (10), has the KL property at* (x^*, x^*)*. Then, for any* $x_0 \in \mathbb{R}^N$*, and* α*,* β *selected according the property of* g *in one of the last three rows of Table 1 with* $L = \lambda^{-1}$*, the sequence* $(x^k)_{k\in\mathbb{N}}$
   - *has the finite length property,*
   - *and converges to a critical point* $\tilde{x}$ *of* h *with* $h(\tilde{x}) = h(x^*)$*.*

If g *is convex, the sequence* $(x^k)_{k\in\mathbb{N}}$ *is uniquely determined.*

Proof The proof follows analogously to that of Proposition 5.3 by, either invoking Proposition 5.2 and Corollary 4.2 or Proposition 5.1 and Corollary 4.1. □

Remark 5.6 Corollary 4.3 provides convergence rates for Proposition 5.5.

Example 5.2 (Inertial alternating projection for the semi-algebraic feasibility problem)

- The algorithm described in Proposition 5.5 can be used to solve the semi-algebraic feasibility problem of Example 3.1 with $M = 2$. The conditions in Example 3.1 are satisfied.
- If S_1 is non-convex and S_2 convex, the second case of Proposition 5.5 yields a *globally convergent relaxed alternating projection method* with $g = \delta_{S_1}$ and $f = \delta_{S_2}$. Table 1 requires the step size conditions $\beta \in [0, \frac{1}{2}[$ and $\alpha \in]0, 1 - 2\beta[$ (note that $\lambda = 1$), which for $\beta = 0$ yields $\alpha \in]0, 1[$, which leads to the following update step:

$$x^{k+1} \in \mathrm{proj}_{S_1}((1-\alpha)x^k + \alpha\, \mathrm{proj}_{S_2}(x^k))$$

Example 5.3 The algorithm described in Proposition 5.5 can be used to solve a relaxed version of the following problem:

$$\min_{x_1,\ldots,x_M \in \mathbb{R}^N} \sum_{i=1}^{M} g_i(x_i)\,, \quad s.t.\; x_1 = \ldots = x_M\,,$$

where the convex constraint is replaced by the associated distance function. The functions $g_i : \mathbb{R}^N \to \overline{\mathbb{R}}$, $i = 1, \ldots, M$, $M \in \mathbb{N}$, are assumed to be proper, lsc, simple, and $x = (x_1, \ldots, x_M) \in \mathbb{R}^{N\times M}$ is the optimization variable. This problem belongs to the second case of Proposition 5.5, i.e., the sequence generated by the inertial alternating proximal minimization method converges globally to a critical point x^* of the function $\sum_{i=1}^{M} g_i(x_i) + \frac{1}{2}(\mathrm{dist}(x, C))^2$ where $C := \{(x_1, \ldots, x_M) \in \mathbb{R}^{N\times M} \;:\; x_1 = \ldots = x_M\}$. The proximal mapping of $\frac{1}{2}(\mathrm{dist}(x, C))^2$ is the projection onto C, which simply averages $x_1, \ldots, x_M$.

5.4 Application: A Feasibility Problem

We consider the example from [54] that demonstrates (local) linear convergence of the alternating projection method. The goal is to find an $N \times M$ matrix X of rank R that satisfies a linear system of equations $\mathcal{A}(X) = B$, i.e.,

$$\text{find} \quad X \quad \text{in} \quad \underbrace{\{X \in \mathbb{R}^{N\times M} \ : \ \mathcal{A}(X) = B\}}_{=:\mathscr{A}} \cap \underbrace{\{X \in \mathbb{R}^{N\times M} \ : \ \operatorname{rank}(X) = R\}}_{=:\mathscr{R}},$$

where $\mathcal{A}\colon \mathbb{R}^{N\times M} \to \mathbb{R}^D$ is a linear mapping and $B \in \mathbb{R}^D$. Such feasibility problems are well suited for split projection methods, as the projection onto each set might be easy to conduct. The projections are given by

$$\operatorname{proj}_{\mathscr{A}}(X) = X - \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}(\mathcal{A}(X) - B) \quad \text{and} \quad \operatorname{proj}_{\mathscr{R}}(X) = \sum_{i=1}^{R} \sigma_i u_i v_i^\top ,$$

where $USV^\top$ is the singular value decomposition of X with $U = (u_1, u_2, \ldots, u_N)$, $V = (v_1, v_2, \ldots, v_M)$ and singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N$ sorted in decreasing order along the diagonal of S. Note that the set of rank-R matrices is a C^2-smooth manifold [55, Ex. 8.14], hence prox-regular [7, Prop. 13.33].

We perform the same experiment as in [54], i.e., we randomly generate an operator $\mathcal{A}$ by constructing random matrices A_1, …, A_D and setting $\mathcal{A}(X) = (\langle A_1, X\rangle, \ldots, \langle A_D, X\rangle)$, $\langle A_i, X\rangle := \operatorname{trace}(A^\top X)$, selecting B such that $\mathcal{A}(X) = B$ has a rank R solution, and the dimensions are chosen as $M = 110$, $N = 100$, $R = 4$, $D = 450$. The performance is measured w.r.t. $|\mathcal{A}(X) - B|$ where X is the result of the projection onto $\mathscr{R}$ in the current iteration.

We consider the *alternating projection* method $X^{k+1} = \operatorname{proj}_{\mathscr{R}}(\operatorname{proj}_{\mathscr{A}}(X^k))$, the *averaged projection* method $X^{k+1} = \frac{1}{2}\big(\operatorname{proj}_{\mathscr{A}}(X^k) + \operatorname{proj}_{\mathscr{R}}(X^k)\big)$, the globally convergent relaxed alternating projection method from Example 5.2 (`glob-altP`, $\alpha = 0.99$), and their inertial variants proposed in Sects. 5.2 and 5.3. For Heavy-ball method/inertial averaged projection (`loc-hb-avgP-bt`, $\beta = 0.75$) in Sect. 5.2 applied to the objective $\operatorname{dist}(X, \mathscr{A})^2 + \operatorname{dist}(X, \mathscr{R})^2$, we use the backtracking line-search version of iPiano [3, Algorithm 4] to estimate the Lipschitz constant automatically. For iPiano/inertial alternating projection (`glob-ipiano-altP`) in Sect. 5.3 applied to $\min_{X\in\mathscr{R}} \frac{1}{2}(\operatorname{dist}(X, \mathscr{A}))^2$ (i.e., g non-convex, f smooth convex), we use $\beta = 0.45 \in [0, \frac{1}{2}[$ and $\alpha = 0.99(1 - 2\beta)/L$ with $L = 1$, which guarantees global convergence to a stationary point, and a backtracking version (`glob-ipiano-altP-bt`) [4, Algorithm 5]. Moreover, for the same setting, we use a heuristic version (`heur-ipiano-altP`, $\beta = 0.75$, theoretically infeasible) with α such that $\alpha\lambda^{-1} = 1$ in Proposition 5.5. Finally, we also consider the locally convergent version of iPiano in Proposition 5.5 (`loc-ipiano-altP-bt`, $\beta = 0.75$) applied to the objective[2] $\min_{X\in\mathscr{A}} \frac{1}{2}(\operatorname{dist}(X, \mathscr{R}))^2$ (i.e., g convex, f prox-regular, non-convex) with backtracking. For the local convergence results, we assume that we start close

[2] The error is measured after projecting the current iterate to the set of rank R matrices.

Table 2 Convergence results for 200 randomly generated feasibility problem as described in Sect. 5.4

Precision $10^p \rightarrow$	− 2	− 4	− 6	− 8	− 10	− 12	− 2	− 4	− 6	− 8	− 10	− 12	− 2	− 4	− 6	− 8	− 10	− 12
Method	Iterations						Time (s)						Success (%)					
`alternating projection`	235	886	–	–	–	–	1.88	7.03	–	–	–	–	100	97.5	0	0	0	0
`averaged projection`	639	–	–	–	–	–	5.13	–	–	–	–	–	100	0	0	0	0	0
`Douglas-Rachford`	974	–	–	–	–	–	8.10	–	–	–	–	–	2	0	0	0	0	0
`Douglas-Rachford 75`	209	449	696	949	–	–	1.68	3.62	5.63	7.66	–	–	100	100	100	100	0	0
`glob-altP`, $\alpha = 0.99$	238	894	–	–	–	–	1.92	7.18	–	–	–	–	100	96.5	0	0	0	0
`glob-ipiano-altP`, $\beta = 0.45$	–	–	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0
`glob-ipiano-altP-bt`, $\beta = 0.45$	45	69	90	115	140	166	0.65	1.03	1.52	2.08	2.63	3.20	100	100	100	100	100	100
`heur-ipiano-altP`, $\beta = 0.75$	59	212	386	567	749	925	0.79	2.82	5.14	7.52	9.93	12.22	100	100	100	100	100	91
`loc-hb-avgP-bt`, $\beta = 0.75$	126	297	502	717	929	–	2.29	5.47	9.24	13.21	17.17	–	100	100	100	100	93.5	0
`loc-ipiano-altP-bt`, $\beta = 0.75$	66	101	138	176	214	252	1.32	2.06	2.80	3.56	4.31	5.06	100	100	100	100	100	100

The table entries show the average number of iterations and the average time that each method requires to reach a certain precision in $\{10^{-2}, 10^{-4}, \ldots, 10^{-12}\}$. A dash ("—") means that the maximum of 1000 iterations was exceeded. The rightmost part of the table lists the success rate of achieving a certain accuracy within 1000 iterations. For a representative example, the convergence is plotted in Fig. 2

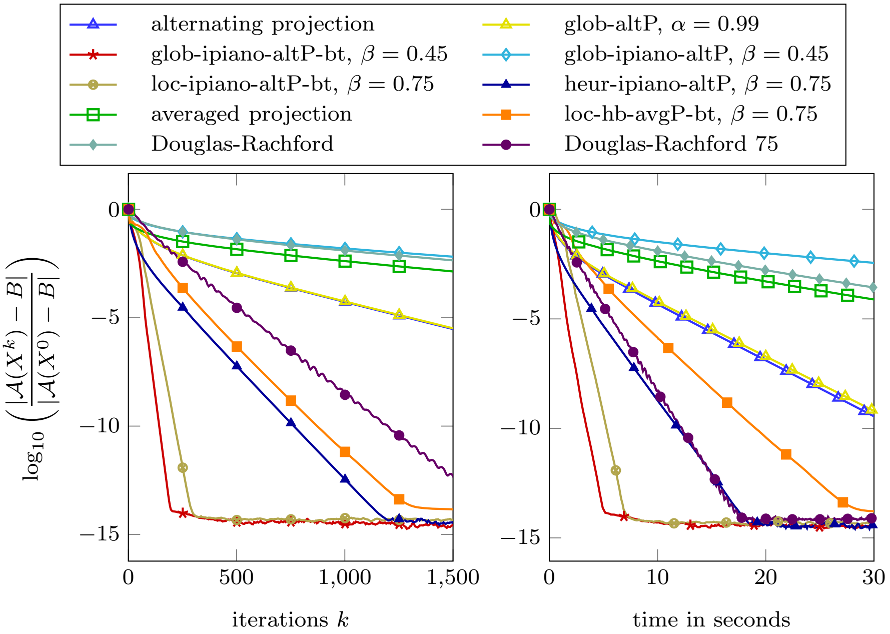

Fig. 2 Convergence plots for the feasibility problem in Sect. 5.4. The inertial methods developed in this paper significantly outperform all other methods with respect to the number of iterations (left plot) and the actual computation time (right plot)

enough to a feasible point. Experimentally, all algorithms converge to a feasible point. In theory, backtracking is not required; however as the radius of the neighborhood of attraction is hard to quantify, the algorithm is more stable with backtracking.

We also compare our method against the recently proposed globally convergent Douglas–Rachford splitting for non-convex feasibility problems [39]. The algorithm depends on a parameter γ, which in theory is required to be rather small: $\gamma_0 := \sqrt{3/2} - 1$. The basic model `Douglas-Rachford` uses the maximal feasible value for this γ-parameter. `Douglas-Rachford 75` is a heuristic version[3] proposed in [39].

Table 2 compares the methods on a set of 200 randomly generated problems with a maximum of 1000 iterations for each method. Also local methods seem to reliably find a feasible point. This seems to be true also for the heuristic methods `Douglas-Rachford 75` and `heur-ipiano-altP`, which shows that there is still a gap between theory and practice. The inertial algorithms that use backtracking significantly outperform methods without backtracking or inertia. Considering the actual computation time makes this observation even more significant, since backtracking algorithms require to compute the objective value several times per iteration.

[3] The heuristic version of Douglas–Rachford splitting in [39] guarantees boundedness of the iterates. We set $\gamma = 150\gamma_0$ and update γ by $\max(\gamma/2, 0.9999\gamma_0)$ if $\|y^k - y^{k-1}\| > t/k$. We refer to [39] for the meaning of y^k. Since the proposed value $t = 1000$ did not work well in our experiment, we optimized t manually. $t = 75$ worked best.

Interestingly, the globally convergent version of iPiano converged the fastest to a feasible point. The convergence behavior of the methods is visualized in Fig. 2 for a representative example.

6 Conclusions

In this paper, we proved a local convergence result for abstract descent methods, which is similar to that of Attouch et al. [8]. This local convergence result is applicable to an inertial forward–backward splitting method, called iPiano [3]. For functions that satisfy the Kurdyka–Łojasiewicz inequality at a local optimum, under a certain growth condition, we verified that the sequence of iterates stays in a neighborhood of a local (or global) minimum and converges to the minimum. As a consequence, the properties that imply convergence of iPiano are required to hold locally only. Combined with a well-known expression for the gradient of Moreau envelopes in terms of the proximal mapping, relations of iPiano to an inertial averaged proximal minimization method and an inertial alternating proximal minimization method are uncovered. These considerations are conducted for functions that are prox-regular instead of the stronger assumption of convexity. For a non-convex feasibility problem, experimentally, iPiano significantly outperforms the alternating projection method and a recently proposed non-convex variant of Douglas–Rachford splitting.

References

1. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
2. Zavriev, S., Kostyuk, F.: Heavy-ball method in nonconvex optimization problems. Comput. Math. Model. **4**(4), 336–341 (1993)
3. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: inertial proximal algorithm for non-convex optimization. SIAM J. Imaging Sci. **7**(2), 1388–1419 (2014)
4. Ochs, P.: Long Term Motion Analysis for Object Level Grouping and Nonsmooth Optimization Methods. Ph.D. thesis, Albert–Ludwigs–Universität Freiburg (2015)
5. Poliquin, R.A., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Am. Math. Soc. **348**(5), 1805–1838 (1996)
6. Poliquin, R.A.: Integration of subdifferentials of nonconvex functions. Nonlinear Anal.: Theory Methods Appl. **17**(4), 385–398 (1991)
7. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer, Berlin (1998)
8. Attouch, H., Bolte, J., Svaiter, B.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**(1–2), 91–129 (2013)
9. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier **48**(3), 769–783 (1998)
10. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. In: Les Équations aux Dérivées Partielles, pp. 87–89. Éditions du centre National de la Recherche Scientifique, Paris (1963)
11. Łojasiewicz, S.: Sur la géométrie semi- et sous- analytique. Annales de l'institut Fourier **43**(5), 1575–1595 (1993)
12. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2006)
13. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)

14. Bochnak, J., Coste, M., Roy, M.F.: Real Algebraic Geometry. Springer, Berlin (1998)
15. Bolte, J., Daniilidis, A., Lewis, A.: A nonsmooth Morse-Sard theorem for subanalytic functions. J. Math. Anal. Appl. **321**(2), 729–740 (2006)
16. den Dries, L.V.: Tame Topology and o-Minimal Structures, London Mathematical Society Lecture Notes Series, vol. 248. Cambridge University Press, Cambridge (1998)
17. Absil, P., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. **16**(2), 531–547 (2005)
18. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
19. Bolte, J., Daniilidis, A., Ley, A., Mazet, L.: Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**, 3319–3363 (2010)
20. Bento, G.C., Soubeyran, A.: A generalized inexact proximal point method for nonsmooth functions that satisfy the Kurdyka–Łojasiewicz inequality. Set-Valued Var. Anal. **23**(3), 501–517 (2015)
21. Noll, D.: Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. J. Optim. Theory Appl. **160**(2), 553–572 (2013)
22. Hosseini, S.: Convergence of nonsmooth descent methods via Kurdyka–Łojasiewicz inequality on Riemannian manifolds. Tech. Rep. 1523, Institut für Numerische Simulation, Rheinische Friedrich–Wilhelms–Universität Bonn, Bonn, Germany (2015)
23. Chouzenoux, E., Pesquet, J.C., Repetti, A.: Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. **162**(1), 107–132 (2014)
24. Bonettini, S., Loris, I., Porta, F., Prato, M., Rebegoldi, S.: On the Convergence of Variable Metric Line-Search Based Proximal-Gradient Method Under the Kurdyka–Lojasiewicz inequality. arXiv:1605.03791 [math] (2016)
25. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. J. Sci. Comput. **72**(2), 700–734 (2017)
26. Chouzenoux, E., Pesquet, J.C., Repetti, A.: A block coordinate variable metric forward–backward algorithm. J. Glob. Optim. **66**(3), 457–485 (2016)
27. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**(3), 874–900 (2014)
28. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka–Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
29. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
30. Bot, R.I., Csetnek, E.R., László, S.: An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. EURO J. Comput. Optim. **4**(1), 3–25 (2015)
31. Ochs, P.: Unifying Abstract Inexact Convergence Theorems for Descent Methods and Block Coordinate Variable Metric iPiano. arXiv:1602.07283 [math] (2016)
32. Bot, R.I., Csetnek, E.R.: An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. J. Optim. Theory Appl. **171**(2), 600–616 (2016)
33. Liang, J., Fadili, J., Peyré, G.: A Multi-step Inertial Forward–Backward Splitting Method for Nonconvex Optimization. arXiv:1606.02118 [math] (2016)
34. Johnstone, P.R., Moulin, P.: Convergence Rates of Inertial Splitting Schemes for Nonconvex Composite Optimization. arXiv:1609.03626v1 [cs, math] (2016)
35. Li, H., Lin, Z.: Accelerated proximal gradient method for nonconvex programming. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NIPS), pp. 379–387 (2015)
36. Stella, L., Themelis, A., Patrinos, P.: Forward–backward quasi-Newton methods for nonsmooth optimization problems. Comput. Optim. Appl. **67**(3), 443–487 (2017)
37. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**(4), 2434–2460 (2015)
38. Li, G., Liu, T., Pong, T.K.: Peaceman-Rachford splitting for a class of nonconvex optimization problems. Comput. Optim. Appl. **68**(2), 407–436 (2017)
39. Li, G., Pong, T.K.: Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Program. **159**(1), 371–401 (2016)

40. Ochs, P., Dosovitskiy, A., Brox, T., Pock, T.: On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. SIAM J. Imaging Sci. **8**(1), 331–372 (2015)
41. Bolte, J., Pauwels, E.: Majorization–minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. Math. Oper. Res. **41**(2), 442–465 (2016)
42. Li, G., Pong, T.: Calculus of the exponent of Kurdyka–Łojasiewicz Inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. (2017). https://doi.org/10.1007/s10208-017-9366-8
43. Merlet, B., Pierre, M.: Convergence to equilibrium for the backward Euler scheme and applications. Commun. Pure Appl. Anal. **9**(3), 685–702 (2010)
44. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)
45. Pock, T., Sabach, S.: Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J. Imaging Sci. **9**(4), 1756–1787 (2016)
46. Poliquin, R., Rockafellar, R., Thibault, L.: Local differentiability of distance functions. Trans. Am. Math. Soc. **352**(11), 5231–5249 (2000)
47. Daniilidis, A., Lewis, A., Malick, J., Sendov, H.: Prox-regularity of spectral functions and spectral sets. J. Convex Anal. **15**(3), 547–560 (2008)
48. Bolte, J., Nguyen, T., Peypouquet, J., Suter, B.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **165**(2), 471–507 (2017)
49. Li, G., Mordukhovich, B., Pham, T.: New fractional error bounds for polynomial systems with applications to Hölderian stability in optimization and spectral theory of tensors. Math. Program. **153**(2), 333–362 (2015)
50. Li, G., Mordukhovich, B., Nghia, T., Pham, T.: Error bounds for parametric polynomial systems with applications to higher-order stability analysis and convergence rates. Math. Program. 1–34 (2016)
51. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, BErlin (2011)
52. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence for alternating and averaged nonconvex projections. Found. Comput. Math. **9**(4), 485–513 (2008)
53. Jourani, A., Thibault, L., Zagrodny, D.: Differential properties of the Moreau envelope. J. Funct. Anal. **266**(3), 1185–1237 (2014)
54. Lewis, A., Malick, J.: Alternating projections on manifolds. Math. Oper. Res. **33**(1), 216–234 (2008)
55. Lee, J.: Introduction to Smooth Manifolds. Graduate Texts in Mathematics 218. Springer, New York (2003)