PAPER • OPEN ACCESS

On an iteratively reweighted linesearch based algorithm for nonconvex composite optimization

View the article online for updates and enhancements.

You may also like

On an iteratively reweighted linesearch based algorithm for nonconvex composite optimization

S Bonettini[1,*] , **D Pezzi[1]** , **M Prato[1]** and **S Rebegoldi[2]**

[1] Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy
[2] Dipartimento di Ingegneria Industriale, Università di Firenze, Viale Morgagni 40, 50143 Firenze, Italy

E-mail: silvia.bonettini@unimore.it

Abstract

In this paper we propose a new algorithm for solving a class of nonsmooth non-convex problems, which is obtained by combining the iteratively reweighted scheme with a finite number of forward–backward iterations based on a linesearch procedure. The new method overcomes some limitations of linesearch forward–backward methods, since it can be applied also to minimize functions containing terms that are both nonsmooth and nonconvex. Moreover, the combined scheme can take advantage of acceleration techniques consisting in suitable selection rules for the algorithm parameters. We develop the convergence analysis of the new method within the framework of the Kurdyka–Łojasiewicz property. Finally, we present the results of a numerical experience on microscopy image super resolution, showing that the performances of our method are comparable or superior to those of other algorithms designed for this specific application.

Keywords: numerical optimization, nonsmooth nonconvex problems, microscopy image super resolution

(Some figures may appear in colour only in the online journal)

---

* Author to whom any correspondence should be addressed.

1. Introduction

Variational models offer a great tool for solving a variety of inverse problems in many fields of application, such as imaging and machine learning, where an explicit direct solution is seldom available. Through a Bayesian argument, they lead to compute the solution of the inverse problem by solving a problem with the following form:

$$\min_{x \in \mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x). \tag{1}$$

Problem (1) is often referred to as a *composite* optimization problem. On the one hand, f_0 represents a fidelity term, which serves the purpose of evaluating the quality of the solution with respect to the data at hand. On the other hand, f_1 is the regularization function, which is related to the prior probability of the unknown, and is used to impose additional information that is believed to be valid in advance. More in general, here and in the following we will simply assume that f_0 contains only smooth terms, coming from the Bayesian paradigm, while all the nonsmooth ones are included in f_1.

Problem (1) becomes especially difficult to handle, from the theoretical and numerical point of view, when either f_0 and /or f_1 are nonconvex. In these cases, the best we can expect from a descent algorithm is the convergence to a stationary point, i.e. a point satisfying the necessary optimality condition, instead of a global or even a local minimizer. The most difficult case in the analysis and implementation of optimization algorithms is when f_1 is nonconvex, i.e. when the objective function contains some terms that are both nonconvex and nonsmooth. When f_1 is convex, that is all the nonsmooth terms are convex, problem (1) can be easily tackled by forward–backward optimization methods [1–5]. The forward step performs a gradient descent on the smooth part of the objective function, and then feeds the result to the backward one, which takes advantage of the convexity of the nonsmooth term to compute the proximal operator, also known as the resolvent. More specifically, the proximal operator associated to f_1 is defined as,

$$\text{prox}^D_{\alpha f_1}(z) = \underset{x \in \mathbb{R}^n}{\text{argmin}} f_1(x) + \frac{1}{2\alpha} \|x - z\|^2_D, \tag{2}$$

where $\alpha > 0$ is the steplength parameter, D is a symmetric and positive definite matrix, and $\|\cdot\|_D$ denotes the norm with respect to the metric induced by the matrix D. A quite general forward–backward scheme can be summed in the following two steps:

$$y^{(k)} = \text{prox}^{D_k}_{\alpha_k f_1}(x^{(k)} - \alpha_k D_k^{-1} \nabla f_0(x^{(k)})) \tag{3}$$

$$x^{(k+1)} = x^{(k)} + \lambda_k(y^{(k)} - x^{(k)}). \tag{4}$$

Many algorithms can be cast in this framework, with each one exploiting the three parameters α_k, D_k and λ_k in different ways to achieve convergence and effectiveness. Besides the standard choices for α_k and λ_k as fixed values related to the Lipschitz constant of ∇f_0 [1, 2], adaptive procedures for computing these steplength parameters can be included. Regarding the steplength selection, we may adopt backtracking (adaptive computation of α_k with $\lambda_k = 1$ [6, 7]), or linesearch approaches (adaptive computation of λ_k for a given $\alpha_k > 0$ [3–5]). As concerns the choice of the matrix D_k, we mention the approach in [3], where the matrix D_k must be selected in order to define a quadratic majorant of the objective function, and the linesearch based approach in [4], where both the matrix D_k and the steplength α_k can be considered as almost entirely free parameters, which can be exploited to improve its practical performances. The convergence analysis of forward–backward methods has also received much attention in the last years. Besides the classical results obtained under convexity assumptions

on both f_0 and f_1 [1, 2], more general results have been obtained by assuming that the objective function f satisfies the Kurdyka–Łojasiewicz property, for several different choices of the algorithms parameters [3, 7–11].

On the other side, the case when f_1 is nonsmooth and nonconvex is more difficult to handle in a forward–backward scheme, from both the theoretical and practical point of view. An easy example of these difficulties is the fact that, when f_1 is nonconvex, the minimization problem underlying the resolvent operator may not have a unique solution. However, variational models including nonsmooth nonconvex terms are receiving an increasing interest, since they allow the modelling of more accurate noise statistics [12, 13] and the recovery of sparse solutions, which are typically promoted by including an ℓ_0 penalization (or a continuous, nonconvex approximation of it) in the objective function [14–16]. The second kind of problem is especially relevant in the field of single-molecule localization microscopy (SMLM) [17–19], where the object represented in the acquired images is typically very sparse. From the theoretical viewpoint, iteration (3) and (4) can be still safely applied also when f_1 is nonconvex and nonsmooth [8, 20, 21]; however, to the best of our knowledge, all the related convergence analyses require that $\lambda_k = 1$ for all k, thus limiting the possibility to adaptively compute the algorithm parameters.

The iteratively reweighted methods, whose idea dates back to [22], have been developed in the last years as a tool to solve (1) when the (possibly) nonsmooth term f_1 can be written as the composition of a nonconvex smooth function with a convex one. The main idea underlying these methods consists in replacing problem (1) with a sequence of convex subproblems, whose objective function is a majorant (or surrogate) of the original one [23]. In this sense, the iteratively reweighted methods are strictly related to majorization–minimization methods [3, 12] and they require the minimization of the surrogate function at each iterate. In the recent literature, this kind of method has been applied to the minimization of the ℓ_p norm with $p < 1$ [24] and, in particular, for the case $p = 0$ [14, 25, 26]. In the last few years, the analysis has been extended also to more general problems, within the framework of the Kurdyka–Łojasiewicz property [23, 27].

1.1. Contribution

In this paper we propose a new algorithm that combines the basic iteratively reweighted scheme with a forward–backward linesearch based procedure. In particular, we focus on problem (1) where:

$$f_1(x) = \sum_{j=1}^{n} (\phi_j \circ \psi_j)(x) + \chi(x), \tag{5}$$

with $\psi_j : \mathbb{R}^n \to \mathbb{R}$ convex and lower semicontinuous, $\phi_j : \mathbb{R} \to \mathbb{R}$ concave, nondecreasing and continuously differentiable, $j = 1, \dots, J$, and $\chi : \mathbb{R}^n \to \mathbb{R}$ convex and lower semicontinuous. The algorithm iteration is derived by a two-step procedure: first, a majorizer for f_1 is obtained by exploiting the concavity of ϕ_j; then, the surrogate problem is approximately solved by a finite, bounded number of steps of a linesearch forward–backward method. We develop the convergence analysis of the new method within the framework of the Kurdyka–Łoyasiewicz property. Even if the assumptions needed for the theoretical analysis, which will be detailed later, prevents to directly handle the ℓ_p norms with $p < 1$, our approach can be still applied to several interesting cases. In particular, we will apply our algorithm to the regularization proposed in [16], named CEL0, as an approximation of the ℓ_0 norm for image super resolution in microscopy. We will show that this functional can be set in the form (5), and we will present the results of a numerical experience in this framework, comparing our proposed algorithm

to other iteratively reweighted methods presented in the literature for this specific application [16, 17, 28].

1.2. Related work

The idea of plugging forward–backward iterations in an outer iteratively reweighted / majorization-minimization scheme is proposed also in [27]. Although the theoretical settings are very similar, in our approach we exploit a quite different forward–backward iteration based on a descent direction and a linesearch procedure along it [4, 9]. This results in a larger freedom of choosing the parameters α_k and D_k, which, in turn, can lead to improved practical performances. In this way we can introduce acceleration techniques through suitable steplength and scaling selection strategies, such as the Barzilai–Borwein rules for α_k [29] and the Split Gradient technique for D_k [30]. Moreover, we generalize the analysis in such a way that the proximity operator of the convex part of the surrogate problem can be computed inexactly according to an implementable inexactness condition, in the same spirit as in [9].

1.3. Paper organization

The outline of the paper is as follows. In section 2 we recall preliminary results regarding subdifferential calculus and the inexact proximal operator computation that are needed later. Section 3 is dedicated to our proposed algorithm: in 3.1 we state the problem of interest under some appropriate assumptions on the involved functions; in 3.2 the proposed iteratively reweighted variable metric inexact linesearch algorithm (IR-VMILA) is reported and explained in detail; in 3.3 we provide an insightful analysis of its convergence properties. Finally, in section 4 we report the results of our numerical testing, with 4.1 being focused on an image deconvolution toy problem, while 4.3 illustrates a more interesting application to super-resolution in microscopy where a nonconvex regularizer is employed. Some concluding remarks are offered in section 5.

2. Preliminaries

2.1. Notations

In this paper we will adopt a discrete setting and the reference space is $\mathbb{R}^n$. The symbol $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$, i.e. $\langle x, y \rangle = x^T y$ for all $x, y \in \mathbb{R}^n$. We denote with $\| \cdot \|$ the Euclidean norm of a vector, i.e. $\|x\| = \sqrt{\langle x, x \rangle}$ for all $x \in \mathbb{R}^n$, whereas the symbol $\| \cdot \|_D$ indicates the norm induced by the symmetric positive definite matrix D, that is $\|x\|_D = \sqrt{\langle x, Dx \rangle}$. The symbol $\mathcal{S}(\mathbb{R}^n)$ denotes the set of all $n \times n$ symmetric matrices, whereas $\mathcal{S}_{++}(\mathbb{R}^n)$ is the set of all $n \times n$ symmetric positive definite matrices. Furthermore, we employ the notation $\mathcal{I}_n \in \mathbb{R}^{n \times n}$ to denote the identity matrix of order n. On the set $\mathcal{S}(\mathbb{R}^n)$, we consider the Loewner partial ordering relation, which is defined as follows:

$$\forall D_1, D_2 \in \mathcal{S}(\mathbb{R}^n), \quad D_1 \preceq D_2 \quad \Leftrightarrow \quad \langle D_1 x, x \rangle \leqslant \langle D_2 x, x \rangle, \ \forall x \in \mathbb{R}^n.$$

We denote by $\bar{\mathbb{R}}$ the extended real line, i.e. $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$. Given a function $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$, the domain of f is denoted as $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$, and f is called proper if $\mathrm{dom}(f) \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. If F is a set valued mapping defined on $\mathbb{R}^n$, its domain is defined as $\mathrm{dom}(F) = \{x \in \mathbb{R}^n : F(x) \neq \emptyset\}$. Given $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ and $l, u \in \mathbb{R}$, with $l \leqslant u$, we adopt also the following notations:

$$[l < f < u] = \{x \in \mathbb{R}^n : l < f(x) < u\}, \quad [f \leqslant u] = \{x \in \mathbb{R}^n : f(x) \leqslant u\}.$$

The distance of a point $x \in \mathbb{R}^n$ from a set $\Omega \subseteq \mathbb{R}^n$ is defined as $\text{dist}(x, \Omega) = \inf_{y \in \Omega} \|x - y\|$. The relative interior of a convex set Ω is the interior of Ω relative to the smallest affine set $\text{aff}(\Omega)$ containing Ω, i.e. $\text{relint}(\Omega) = \{x \in \Omega : \exists\, \epsilon > 0 \text{ s.t. } B(x, \epsilon) \cap \text{aff}(\Omega) \subseteq \Omega\}$. Finally, given a set $A \subseteq \mathbb{R}^n$, the characteristic function $\mathbb{1}_A$ is such that $\mathbb{1}_A(x) = 1$ if $x \in A$, and $\mathbb{1}_A(x) = 0$ otherwise.

2.2. Subdifferential calculus

In this section we summarize some basic results concerning the subdifferential calculus of (possibly nonconvex) functions. Most of the definitions and results reported here are from [31, 32].

We start by reporting the definitions of regular vector and regular function. The latter concept is of paramount importance for deriving the subdifferential calculus rules required for our convergence analysis.

Definition 1. [31, definition 6.3] Let $\Omega \subseteq \mathbb{R}^n$ and $x \in \Omega$. A vector $v \in \mathbb{R}^n$ is regular normal to Ω at x, i.e. $v \in \widehat{N}_\Omega(x)$, if:

$$\limsup_{y \xrightarrow{\Omega} x,\ y \neq x} \frac{\langle v, y - x \rangle}{\|y - x\|} \leqslant 0,$$

where $y \xrightarrow{\Omega} x$ means 'for all sequences $\{y^{(k)}\}_{k \in \mathbb{N}} \subseteq \Omega$ converging to x'.

A vector $v \in \mathbb{R}^n$ is normal to Ω at x, i.e. $v \in N_\Omega(x)$, if there exist sequences $\{y^{(k)}\}_{k \in \mathbb{N}} \subseteq \Omega$, $\{v^{(k)}\} \subseteq \mathbb{R}^n$ with $v^{(k)} \in \hat{N}_\Omega(y^{(k)})$ such that $y^{(k)} \to x$ and $v^{(k)} \to v$.

Definition 2. [31, definition 7.25] A function $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is called (Clarke) regular at x if $f(x)$ is finite and the epigraph $\text{epi}(f) := \{(x, t) : x \in \text{dom}(f),\ t \geqslant f(x)\}$ is Clarke regular at $(x, f(x))$ as a subset of $\mathbb{R}^n \times \mathbb{R}$, namely, if $\text{epi}(f)$ is locally closed and it holds $N_{\text{epi}(f)}(x) = \widehat{N}_{\text{epi}(f)}(x)$.

Next, we recall the definitions of Fréchet subdifferential, limiting-subdifferential, and horizon-subdifferential.

Definition 3. [31, definition 8.3] Let $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ and $x \in \text{dom}(f)$. The Fréchet subdifferential of f at x is the set:

$$\hat{\partial}f(x) = \left\{ v \in \mathbb{R}^n : \liminf_{y \to x, y \neq x} \frac{1}{\|x - y\|} (f(y) - f(x) - (y - x)^T v) \geqslant 0 \right\}.$$

The limiting-subdifferential (or simply subdifferential) of f at x is defined as:

$$\partial f(x) = \{v \in \mathbb{R}^n : \exists\, \{y^{(k)}\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^n,\ v^{(k)} \in \hat{\partial}f(y^{(k)})\ \forall k \in \mathbb{N} \text{ such that }$$
$$y^{(k)} \to x,\ f(y^{(k)}) \to f(x) \text{ and } v^{(k)} \to v\}.$$

Finally, the horizon-subdifferential of f at x is given by:

$$\partial^\infty f(x) = \{v \in \mathbb{R}^n : \exists\, \{y^{(k)}\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^n,\ v^{(k)} \in \hat{\partial}f(y^{(k)}),\ \lambda^{(k)} > 0\ \forall k \in \mathbb{N} \text{ such that }$$
$$y^{(k)} \to x,\ f(y^{(k)}) \to f(x),\ \lambda^{(k)} \to 0 \text{ and } \lambda^{(k)} v^{(k)} \to v\}.$$

Definition 4. Let $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$. A point $x \in \mathbb{R}^n$ is stationary for f if $x \in \text{dom}(f)$ and $0 \in \partial f(x)$.

Remark 1. Given $f \colon \mathbb{R}^n \to \bar{\mathbb{R}}$, we have that any (local) minimum point $x \in \mathbb{R}^n$ of f is a stationary point, while the converse may not be true in general. If f is also convex, then x is a (local) minimum point if and only if x is a stationary point.

In the following lemma, we state a useful closedness property satisfied by the limiting sub-differential. Such a property will be employed to show that the limit points of our proposed algorithm are stationary, as similarly done in other works in the literature [3, 27].

Lemma 1. *[31, theorem 8.6] Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ and $x \in \mathrm{dom}(f)$. Suppose $\{(x^{(k)}, v^{(k)})\}_{k \in \mathbb{N}}$ is a sequence such that $v^{(k)} \in \partial f(x^{(k)})$ for all $k \geqslant 0$. If $\lim_{k \to \infty} x^{(k)} = x$, $\lim_{k \to \infty} v^{(k)} = v$, and $\lim_{k \to \infty} f(x^{(k)}) = f(x)$, then $v \in \partial f(x)$.*

We now report some basic subdifferential calculus rules under appropriate convexity or regularity assumptions on the involved functions.

Lemma 2. *(i) [31, exercise 8.8c] If $f = f_0 + f_1$ with f_1 finite at x and f_0 continuously differentiable on a neighbourhood of x, then:*

$$\partial f(x) = \nabla f_0(x) + \partial f_1(x).$$

(ii) [32, p 222] Let $\psi : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, convex, lower semicontinuous, and $\lambda \geqslant 0$. If $x \in \mathrm{dom}(\partial \psi)$, then:

$$\partial(\lambda \psi)(x) = \lambda \partial \psi(x).$$

(iii) [32, theorem 23.8] Let $\psi_j : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, convex, and lower semicontinuous for all $j = 1, \dots, J$, and define $\psi = \psi_1 + \dots + \psi_J$. If $\mathrm{relint}(\mathrm{dom}(\psi_i)) \cap \mathrm{relint}(\mathrm{dom}(\psi_j)) \neq \emptyset$ for all $i, j = 1, \dots, J$, then for all $x \in \bigcap_{j=1}^{J} \mathrm{dom}(\partial \psi_j)$ there holds:

$$\partial \psi(x) = \sum_{j=1}^{J} \partial \psi_j(x).$$

(iv) [31, corollary 10.9] Let $f_j : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, lower semicontinuous, for all $j = 1, \dots, J$, define $f = f_1 + \dots + f_J$, and let $x \in \mathrm{dom}(f)$. Suppose that the only combinations of subgradients $v_j \in \partial^\infty f_j(x)$ with $v_1 + \dots + v_J = 0$ is $v_1 = \dots = v_J = 0$, and that each f_j is regular at x. Then, also f is regular at x, and

$$\partial f(x) = \sum_{j=1}^{J} \partial f_j(x).$$

The next lemma provides the chain rule for functions of the form (5) and will be employed in the subsequent analysis. The rule can be derived by collecting and putting together some known results of variational analysis contained in [31]. Similar results have been obtained also in [23, 27] under slightly different conditions on the functions of interest.

Lemma 3. *Let $\Psi : \mathbb{R}^n \to \mathbb{R}^J$, $\Psi(x) = (\psi_1(x), \dots, \psi_J(x))$ for all $x \in \mathbb{R}^n$, where the functions $\psi_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \dots, J$, are convex and lower semicontinuous. Furthermore, let $\phi : \mathbb{R}^J \to \bar{\mathbb{R}}$, $\phi(z) = \sum_{j=1}^{J} \phi_j(z_j)$ for all $z = (z_1, \dots, z_J) \in \mathbb{R}^J$, where the functions ϕ_j, $j = 1, \dots, J$, are monotone nondecreasing, and continuously differentiable on an open set Ω_j with $\psi_j(\mathbb{R}^n) \subseteq \Omega_j \subseteq \mathbb{R}$. Finally, let $\chi : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, convex, lower semicontinuous, and locally Lipschitz continuous on its domain. Then, for all $x \in \mathrm{dom}(\partial \chi)$, there holds:*

$$\partial(\phi \circ \Psi + \chi)(x) = \sum_{j=1}^{J} \phi_j'(\psi_j(x)) \partial \psi_j(x) + \partial \chi(x). \tag{6}$$

Proof. First, we focus on the computation of $\partial(\phi \circ \Psi)(x)$. To this aim, we show that the assumptions of the chain rule provided in [31, theorem 10.49] are satisfied. In particular, we observe that: (i) ϕ is Clarke regular at $\Psi(x)$, as it is given by the sum of J separable and Clarke regular functions [31, proposition 10.5]: indeed, each function ϕ_j, $j = 1, \ldots, J$ is Clarke regular at $\psi_j(x)$, being continuously differentiable on an open set Ω_j containing $\psi_j(x)$ [31, example 7.28]; (ii) $\partial^\infty \phi(\Psi(x)) = \{0\}$: indeed, since Ψ is the sum of J separable and Clarke regular functions, it follows from [31, proposition 10.5] that $\partial^\infty \phi(\Psi(x)) = \prod_{j=1}^J \partial^\infty \phi_j(\psi_j(x))$; now, we have $\partial^\infty \phi_j(\psi_j(x)) = \{0\}$ for all $j = 1, \ldots, J$, again because ϕ_j is continuously differentiable on an open set containing $\psi_j(x)$ [31, exercise 8.8(b)]; (iii) the function $s(z) = \langle \nabla \phi(\Psi(x)), \Psi(z) \rangle = \sum_{j=1}^J \phi_j'(\psi_j(x))\psi_j(z)$ is Clarke regular at x; indeed, since each ψ_j is convex and lower semicontinuous, and $\phi_j'(\psi_j(x)) \geqslant 0$ due to the monotonicity of ϕ_j, it turns out that also s is convex and lower semicontinuous, which implies that s is regular [31, example 7.27].

Hence, since properties (i)–(iii) are satisfied, we can apply the chain rule provided in [31, theorem 10.49], and thus conclude that $\phi \circ \Psi$ is regular at x and the following equations hold,

$$\partial(\Phi \circ \Psi)(x) = \partial\left(\langle \nabla\phi(\Psi(x)), \Psi \rangle\right)(x) = \sum_{j=1}^J \phi_j'(\psi_j(x))\partial\psi_j(x), \tag{7}$$

where the first equation is due to the application of [31, theorem 10.49], whereas the second one follows from lemma 2(ii) and (iii), which is applicable due to the fact that each ψ_j is convex and lower semicontinuous, and $\phi_j'(\psi_j(x)) \geqslant 0$.

Now, we turn to the computation of $\partial(\phi \circ \Psi + \chi)(x)$. First, observe that $\phi \circ \Psi$ is locally Lipschitz continuous at x, since it is the composition of two locally Lipschitz functions at x, as ϕ is continuously differentiable and Ψ is convex and finite [31, example 9.14]. Second, χ is locally Lipschitz continuous on its domain by assumption. Then, we can apply [31, theorem 9.13] and deduce that $\partial^\infty(\phi \circ \Psi) = \partial^\infty \chi(x) = \{0\}$. Furthermore, both $\phi \circ \Psi$ and χ are regular functions at x. Then, we can employ lemma 2(iv) with $J = 2, f_1 = \phi \circ \Psi, f_2 = \chi$, together with (7), to get the thesis. $\qquad \square$

As the objective function in (1) may be nonconvex, we need to replace convexity with a different analytical property that still allows us to ensure convergence of the proposed method. Several works in the past literature have compensated for the lack of convexity by assuming that the objective function satisfies the so-called Kurdyka–Łojasiewicz inequality [8, 20, 33].

Definition 5. [20, definition 3] Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, lower semicontinuous. The function f satisfies the Kurdyka–Łojasiewicz (KL) inequality at the point $x^* \in \text{dom}(\partial f)$ if there exist $\nu > 0$, a neighbourhood U of x^*, and a continuous concave function $\xi : [0, \nu) \to [0, +\infty)$ such that $\xi(0) = 0, \xi$ is C^1 on $(0, \nu), \xi'(s) > 0$ for all $s \in (0, \nu)$, and the following inequality holds,

$$\xi'(f(x) - f(x^*))\text{dist}(0, \partial f(x)) \geqslant 1, \ \forall x \in U \cap [f(x^*) < f < f(x^*) + \nu]. \tag{8}$$

If f satisfies the KL inequality for all $x^* \in \text{dom}(\partial f)$, then f is called a KL function.

The KL inequality is satisfied on the entire domain by indicator functions of semi-algebraic sets, real polynomials, $p-$norms and, in general, semi-algebraic functions or real analytic functions (see [20] and references therein for more details). In practice, most objective functions arising in imaging inverse problems do comply with the KL assumption.

In the upcoming convergence analysis, it will be convenient to employ the following uniformized version of the KL inequality, which holds whenever the function satisfies (8) over a compact set upon which it is constant.

Lemma 4. *[20, lemma 6] Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper, lower semicontinuous function and $\Omega^* \subseteq \mathbb{R}^n$ a compact set. Suppose that f satisfies the KL inequality for each point of Ω^*, and that f is constant over Ω^*, i.e. $f(x^*) = f^*$ for all $x^* \in \Omega^*$. Then, there exist $\rho, \nu > 0$ and a function $\xi : [0, \nu) \to [0, +\infty)$ satisfying the same properties as in definition 5 such that:*

$$\xi'(f(x) - f^*)\mathrm{dist}(0, \partial f(x)) \geqslant 1, \quad \forall\, x \in B^*, \tag{9}$$

where the set $B^* \subseteq \mathbb{R}^n$ is given by:

$$B^* = \{x \in \mathbb{R}^n : \mathrm{dist}(x, X) \leqslant \rho \text{ and } f^* < f(x) < f^* + \nu\}. \tag{10}$$

2.3. Inexact computation of a proximal–gradient point

The aim of this section is to provide the settings for the forward–backward iterations employed in our iteratively reweighted scheme. In particular, we will include the possibility of inexactly computing the proximal operator in the backward step, according to an implementable inexactness criterion. To this end, we follow the discussion in [9, section 3], reporting some useful results that will be needed in the next section.

Let $f_0 : \Omega_0 \to \mathbb{R}$ and $\ell : \mathbb{R}^n \to \bar{\mathbb{R}}$, where f_0 is continuously differentiable with $L-$Lipschitz continuous gradient on the open set $\Omega_0 \subseteq \mathbb{R}^n$ with $\overline{\mathrm{dom}(\ell)} \subseteq \Omega_0$, while ℓ is proper, convex, and lower semicontinuous. Given $x \in \mathrm{dom}(\ell)$, we introduce the function:

$$h_\gamma(z; x) = \langle \nabla f_0(x), z - x \rangle + \frac{\gamma}{2\alpha}\|z - x\|_D^2 + \ell(z) - \ell(x), \quad \forall\, z \in \mathbb{R}^n, \tag{11}$$

where $\gamma \in [0, 1]$, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ being $0 < \alpha_{\min} \leqslant \alpha_{\max}$, and $D \in \mathbb{R}^{n \times n}$ such that $\frac{1}{\mu}\mathcal{I}_n \preceq D \preceq \mu\mathcal{I}_n$. Denoting $h(\cdot; \cdot) := h_1(\cdot; \cdot)$, we observe that:

$$h_\gamma(z; x) \leqslant h(z; x), \quad \forall\, z \in \mathbb{R}^n,\ x \in \mathrm{dom}(\ell). \tag{12}$$

Thanks to the assumptions on α and D, we also have that $h(\cdot; x)$ is a strongly convex function with modulus $1/(\alpha_{\max}\mu)$, that is:

$$h(z; x) \geqslant h(y; x) + \langle w, z - y \rangle + \frac{1}{2\alpha_{\max}\mu}\|z - y\|^2, \quad \forall\, y, z \in \mathrm{dom}(\ell), \forall\, w \in \partial h(y; x).$$

Therefore, $h(\cdot; x)$ admits a unique minimum point $\hat{y} \in \mathbb{R}^n$, which we call the proximal–gradient point. Indeed, using the definition of proximal operator in (2), remark 1, and lemma 2(i), we easily deduce that $\hat{y}$ is the point obtained by applying a proximal–gradient step to the function $f_0 + \ell$ at the point x, namely:

$$\hat{y} = \underset{z \in \mathbb{R}^n}{\mathrm{argmin}}\, h(z; x) = \mathrm{prox}_{\alpha\ell}^D(x - \alpha D^{-1}\nabla f_0(x)). \tag{13}$$

We are now interested in the following notion of inexact proximal–gradient point.

Definition 6. [9, definition 3] *Given $x \in \mathrm{dom}(\ell)$, $\tau \geqslant 0$, we call inexact proximal–gradient point any approximation $\tilde{y} \in \mathrm{dom}(\ell)$ of the proximal–gradient point $\hat{y}$ such that,*

$$h(\tilde{y}; x) - h(\hat{y}; x) \leqslant -\frac{\tau}{2}h(\tilde{y}; x). \tag{14}$$

Furthermore, we denote any such approximation of the proximal–gradient point with the notation $\tilde{y} \approx \hat{y}$.

Remark 2. If $\tau = 0$, then the inexactness criterion (14) becomes $h(\tilde{y};x) = h(\hat{y};x)$, and thus $\tilde{y}$ reduces to the exact proximal–gradient point, i.e. $\tilde{y} = \hat{y}$.

Remark 3. Note that the following inequalities hold true:

$$h_\gamma(\tilde{y};x) \leqslant h(\tilde{y};x) \leqslant \left(\frac{2}{2+\tau}\right) h(\hat{y};x) \leqslant \left(\frac{2}{2+\tau}\right) h(x;x) = 0,$$

where the first inequality comes from the definition of $h_\gamma(\cdot;x)$ in (11) and $\gamma \in [0,1]$, the second one from the inexactness condition (14), the third one from the fact that $\hat{y}$ is the minimum point of $h(\cdot;x)$, whereas the last equality is due to the definition of $h(\cdot;x)$. Therefore, we can conclude that:

$$h_\gamma(\tilde{y};x) \leqslant 0.$$

The inexactness criterion (14) has been investigated and employed in several works (see e.g. [4, 5, 9, 34]). Unlike other conditions available in the literature [3, 8, 27], it has the advantage to be practically implementable in certain cases of interest. In particular, if ℓ is the composition of a linear operator with a convex function, then we can compute a point $\tilde{y}$ satisfying (14) by first applying an iterative method to the dual problem of (13), and then stopping the iterates when the primal–dual gap satisfies an appropriate stopping condition (see [9, section 3.1] for further details on this practical procedure).

To conclude this section, we recall some inequalities that involve the points $\tilde{y}$, $\hat{y}$, and x.

Lemma 5. *Let $x \in \mathrm{dom}(\ell)$, and $\tilde{y}$ defined as in (14). Then, the following inequalities hold:*

$$\frac{1}{2}\|\hat{y} - x\|^2 \leqslant \alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)(-h(\tilde{y};x)) \tag{15}$$

$$\frac{1}{2}\|\tilde{y} - \hat{y}\|^2 \leqslant \frac{\alpha_{\max}\mu\tau}{2}(-h(\tilde{y};x)) \tag{16}$$

$$\frac{1}{4}\|\tilde{y} - x\|^2 \leqslant \alpha_{\max}\mu(1+\tau)(-h(\tilde{y};x)). \tag{17}$$

Proof. Inequalities (15) and (16) follow by putting together the strong convexity of $h(\cdot;x)$ with the inexactness condition (14) and the assumptions on α and D. Inequality (17) is obtained by combining the first two inequalities through the triangular inequality and the basic relation $2\sqrt{uv} \leqslant u + v$ (see [9, lemma 2] for the full proof). □

Lemma 6. *Let $x \in \mathrm{dom}(\ell)$, and $\hat{y}, \tilde{y}$ defined as in (13) and (14), respectively. Then, there exists $c \geqslant 0$ such that the following inequalities hold:*

$$(f_0 + \ell)(\hat{y}) \geqslant (f_0 + \ell)(\tilde{y}) - c(-h(\tilde{y};x)), \quad c > 0 \tag{18}$$

$$(f_0 + \ell)(\tilde{y}) \leqslant (f_0 + \ell)(x) - 2L\alpha_{\max}\mu(1+\tau)h(\tilde{y};x) \tag{19}$$

$$(f_0 + \ell)(\hat{y}) \leqslant (f_0 + \ell)(x) - L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)h(\tilde{y};x). \tag{20}$$

Proof. The inequalities (18)–(20) can be obtained by applying [9, lemmas 4 and 5] with $f_1 = \ell$ and $f = f_0 + f_1$. □

3. Proposed algorithm and convergence analysis

3.1. Problem formulation

From now on, we address the following minimization problem:

$$\underset{x\in\mathbb{R}^n}{\operatorname{argmin}} f(x) \equiv f_0(x) + f_1(x), \tag{21}$$

and make the following assumptions on the functions f_0, f_1.

Assumption 1.

(i) $f_0 : \Omega_0 \to \mathbb{R}$ *is continuously differentiable with L-Lipschitz continuous gradient on the open set* $\Omega_0 \subseteq \mathbb{R}^n$, *and* $\overline{\operatorname{dom}(f_1)} \subseteq \Omega_0$.

(ii) $f_1 : \mathbb{R}^n \to \bar{\mathbb{R}}$ *has the form:*

$$f_1(x) = \sum_{j=1}^{J} (\phi_j \circ \psi_j)(x) + \chi(x), \quad \forall\, x \in \mathbb{R}^n, \tag{22}$$

where

. $\psi_j : \mathbb{R}^n \to \mathbb{R}$ *is convex and lower semicontinuous, and* $\phi_j : \mathbb{R} \to \bar{\mathbb{R}}$ *is concave, monotone nondecreasing, continuously differentiable on an open set* Ω_j *with* $\psi_j(\mathbb{R}^n) \subseteq \Omega_j \subseteq \mathbb{R}$, *and with* ϕ_j' *locally Lipschitz continuous on* Ω_j, *for all* $j = 1,\ldots,J$;

. $\chi : \mathbb{R}^n \to \bar{\mathbb{R}}$ *is proper, convex, lower semicontinuous, and locally Lipschitz continuous on its domain;*

(iii) *The function f is coercive, that is* $\lim_{\|x\|\to+\infty} f(x) = +\infty$.

Remark 4. Let us comment on the properties of f_1 required in assumption 1.

(i) The general model (22) is especially suited for nonconvex nonsmooth regularization arising in imaging problems. On the one hand, the term $\sum_{j=1}^{J}(\phi_j \circ \psi_j)(x)$ encompasses several nonconvex regularizers employed in the literature, such as the log-sum penalty term, the Cauchy penalty term, and smoothed versions of the ℓ_p-norm with $p < 1$ (see [23, 27] and our section 4 for other examples). On the other hand, the function χ may represent an easy constraint on x. For instance, we may enforce nonnegativity in the solution by setting $\chi = \iota_{\geqslant 0}$, where $\iota_{\geqslant 0}$ denotes the indicator function of the nonnegative orthant

$$\iota_{\geqslant 0}(x) = \begin{cases} 0 & \text{if } x \geqslant 0 \\ +\infty & \text{otherwise .} \end{cases}$$

This constraint is especially relevant in imaging applications.

(ii) Since the functions ψ_j, $j = 1,\ldots,J$, are convex and finite on $\mathbb{R}^n$, it follows that they are locally Lipschitz continuous [31, example 9.14]. Hence, each subdifferential $\partial\psi_j$ is locally bounded [31, theorem 9.16]. Therefore, for any compact set $K \subseteq \mathbb{R}^n$ there must exist $\xi_K > 0$ such that, for all $x \in K$, $j = 1,\ldots,J$, and $u_j(x) \in \partial\psi_j(x)$, there holds $\|u_j(x)\| \leqslant \xi_K$.

(iii) Being ϕ_j' and ψ_j both locally Lipschitz continuous, we have that $\phi_j' \circ \psi_j$ is Lipschitz continuous on every compact subset of $\mathbb{R}^n$. This means that, for any compact set $K \subseteq \mathbb{R}^n$, there must exist $\rho_K > 0$ such that $\|\phi_j' \circ \psi_j(x) - \phi_j' \circ \psi_j(y)\| \leqslant \rho_K\|x - y\|$ for all $x, y \in K$ and $j = 1,\ldots,J$.

For our convergence analysis, we will require an additional assumption on the objective function, namely that it satisfies the KL inequality on its domain. Such an assumption will be needed only for the final convergence result, hence we state it later in the next section.

3.2. The algorithm

In this section, we introduce our proposed method to solve problem (21), which is denoted as IR-VMILA.

At each outer iteration $k \geqslant 0$, IR-VMILA defines a convex majorant $q^{(k)}$ for the (possibly nonconvex) function f_1, which is obtained by linearizing the functions ϕ_j, $j = 1, \ldots, J$ around the current iterate $x^{(k)}$. Then, the next iterate $x^{(k+1)}$ is computed through the inner iterates $\tilde{x}^{(k,0)}, \tilde{x}^{(k,1)}, \ldots, \tilde{x}^{(k,I_k)}$, which are generated by applying a (uniformly bounded) number of inner iterations I_k of a variable metric linesearch based algorithm to the minimization of the function $f_0(x) + q^{(k)}(x)$.

In order to better describe the steps of IR-VMILA, we introduce some further notations. For all $k \geqslant 0$ and $j = 1, \ldots, J$, we define the following function:

$$q_j^{(k)}(x) := (\phi_j \circ \psi_j)(x^{(k)}) + (\phi_j' \circ \psi_j)(x^{(k)})(\psi_j(x) - \psi_j(x^{(k)})). \tag{23}$$

Since each ϕ_j is concave, it follows that $q_j^{(k)}$ is a majorant of $\phi_j \circ \psi_j$ at $x^{(k)}$, namely

$$\phi_j \circ \psi_j(x) \leqslant q_j^{(k)}(x), \quad \forall x \in \mathbb{R}^n$$
$$\phi_j \circ \psi_j(x^{(k)}) = q_j^{(k)}(x^{(k)}).$$

As a result, for all $k \geqslant 0$, the following function:

$$q^{(k)}(x) := \sum_{j=1}^{J} q_j^{(k)}(x) + \chi(x), \tag{24}$$

is a majorant of f_1 at point $x^{(k)}$. For the sake of simplicity, it will be convenient to consider also the function:

$$\ell^{(k)}(x) = \sum_{j=1}^{J} \omega_j^{(k)} \psi_j(x) + \chi(x), \text{ where } \omega_j^{(k)} = (\phi_j' \circ \psi_j)(x^{(k)}). \tag{25}$$

Note that $\ell^{(k)}$ differs from $q^{(k)}$ by an additive constant, i.e. $\ell^{(k)}(x) = q^{(k)}(x) + c^{(k)}$ with $c^{(k)} \in \mathbb{R}$. In the context of iteratively reweighted algorithms, the parameters $\omega_j^{(k)}, j = 1, \ldots, J$, are usually called *weights*. For all $k \geqslant 0$ and $i = 0, 1, \ldots, I_k - 1$, given $\alpha_{k,i} > 0$ and $D_{k,i} \in \mathcal{S}_{++}(\mathbb{R}^n)$, we define the following functions:

$$h_\gamma^{(k,i)}(z) := \langle \nabla f_0(\tilde{x}^{(k,i)}), z - \tilde{x}^{(k,i)} \rangle + \frac{\gamma}{2\alpha_{k,i}} \|z - \tilde{x}^{(k,i)}\|_{D_{k,i}}^2 + \ell^{(k)}(z) - \ell^{(k)}(\tilde{x}^{(k,i)})$$
$$h^{(k,i)}(z) := h_1^{(k,i)}(z). \tag{26}$$

Finally, the unique minimum point of $h^{(k,i)}(\cdot)$ will be denoted as:

$$\hat{y}^{(k,i)} = \text{prox}_{\alpha_{k,i}\ell^{(k)}}^{D_{k,i}}(\tilde{x}^{(k,i)} - \alpha_{k,i}D_{k,i}^{-1}\nabla f_0(\tilde{x}^{(k,i)})) = \underset{z \in \mathbb{R}^n}{\text{argmin}}\, h^{(k,i)}(z). \tag{27}$$

We are now ready to detail the proposed method IR-VMILA in algorithm 1. At each outer iteration $k \geqslant 0$, we first define the majorant function $q^{(k)}$ of f_1 around the current iterate $x^{(k)}$, by updating the weights $\omega_j^{(k)}$ according to (25) (STEP 1). Then, we initialize the inner routine for inexactly solving the kth subproblem $\min_{x \in \mathbb{R}^n} f_0(x) + q^{(k)}(x)$ by setting $\tilde{x}^{(k,0)} = x^{(k)}$ (STEP 2). Next, the inner routine starts. At each inner iteration $i = 0, 1 \ldots, I_k - 1$, we select a steplength parameter $\alpha_{k,i} > 0$ and a scaling matrix $D_{k,i} \in \mathcal{S}_{++}(\mathbb{R}^n)$ (STEP 3.1), and we compute

a (possibly inexact) proximal–gradient point $\tilde{y}^{(k,i)} \in \mathbb{R}^n$ (STEP 3.2) according to the following condition

$$h^{(k,i)}(\tilde{y}^{(k,i)}) - h^{(k,i)}(\hat{y}^{(k,i)}) \leqslant -\frac{\tau}{2} h^{(k,i)}(\tilde{y}^{(k,i)}), \tag{28}$$

where $\tau \geqslant 0$. If $\tau = 0$, the proximal–gradient point is computed exactly, i.e. we have $\tilde{y}^{(k,i)} = \hat{y}^{(k,i)}$; otherwise, if $\tau > 0$, we look for an approximation $\tilde{y}^{(k,i)}$ of the exact point $\hat{y}^{(k,i)}$. Based on remark 3, we observe that the point $\tilde{y}^{(k,i)}$ makes the function $h_\gamma^{(k,i)}$ negative or null, namely:

$$h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}) \leqslant 0, \quad \forall\, k \geqslant 0,\ i = 0, 1, \ldots, I_k - 1. \tag{29}$$

This property is crucial for the well-posedness of the next inner steps. Indeed, once $\tilde{y}^{(k,i)}$ is computed, we set $d^{(k,i)} = \tilde{y}^{(k,i)} - \tilde{x}^{(k,i)}$ as the search direction (STEP 3.3) and perform a linesearch procedure for the function $f_0 + \ell^{(k)}$ along $d^{(k,i)}$ (STEP 3.4). More precisely, we compute the linesearch parameter $\lambda_{k,i} = \delta^{m_{k,i}}$, where $\delta \in (0,1)$ and $m_{k,i}$ is the smallest nonnegative integer such that the following Armijo–like condition holds:

$$(f_0 + \ell^{(k)})(\tilde{x}^{(k,i)} + \delta^{m_{k,i}} d^{(k,i)}) \leqslant (f_0 + \ell^{(k)})(\tilde{x}^{(k,i)}) + \beta \delta^{m_{k,i}} h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}). \tag{30}$$

Such a condition has been already considered in the nonsmooth framework for other algorithms [4, 5, 35]. Note that, if $h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}) = 0$, i.e. $\tilde{y}^{(k,i)}$ is the exact proximal gradient point, then condition (30) trivially holds with $m_{k,i} = 0$; otherwise, if $h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}) < 0$, it can be shown that $d^{(k,i)}$ is a descent direction for $f_0 + \ell^{(k)}$ at point $\tilde{x}^{(k,i)}$, and the linesearch procedure based on (30) terminates in a finite number of steps [35, proposition 3.1]. Finally, the next inner iterate is computed as the convex combination $\tilde{x}^{(k,i+1)} = \tilde{x}^{(k,i)} + \lambda_{k,i} d^{(k,i)}$ (STEP 3.5). Once the inner routine has ended, we compute the next outer iterate $x^{(k+1)}$ as the point that retains the smallest objective function value between $\tilde{x}^{(k,I_k)}$ and $\tilde{y}^{(k,0)}$ (STEP 4).

Note that, if we set $J = 1$, $\phi_1(x) = x$, $\psi_1 \equiv 0$ in (22), and $I_k = 1$ for all $k \geqslant 0$ in algorithm 1, then our proposed IR-VMILA reduces to the so-called VMILAn algorithm [4, 9, 36] applied to the minimization of the function $f = f_0 + f_1$, where $f_1 = \chi$ is convex. In other words, we can see IR-VMILA as a generalization of the VMILAn algorithm to completely nonconvex optimization problems, where both f_0 and f_1 are possibly nonconvex.

We also remark that the proposed IR-VMILA has a similar structure to the C2FB algorithm considered in [27], which is also an iteratively reweighted algorithm with an inner forward–backward routine. However, in [27], the steplength parameters are forced to be smaller than one, and the scaling matrices are required to satisfy an appropriate majorization–minimization condition. On the contrary, our proposed algorithm allows for a greater flexibility, as the parameters $\alpha_{k,i}$ and $D_{k,i}$ can be freely chosen according to any rule of preference, provided that they both belong to compact sets (see the forthcoming assumption 2). Furthermore, IR-VMILA adopts a different inexactness criterion for the computation of the proximal point $\tilde{y}^{(k,i)}$, which has the advantage of being practically implementable (see section 2.3).

Algorithm 1. IR-VMILA

$x^{(0)} \in \text{dom}(f_1)$, $\beta, \delta \in (0,1)$, $\gamma \in [0,1]$, $\tau \geqslant 0$.

FOR $k = 0, 1, \ldots$
    STEP 1 Set $\omega_j^{(k)} = (\phi_j' \circ \psi_j)(x^{(k)})$, $j = 1, \ldots, J$, and $\ell^{(k)} = \sum_{j=1}^{J} \omega_j^{(k)} \psi_j + \chi$.
    STEP 2 Set $\tilde{x}^{(k,0)} = x^{(k)}$ and choose $I_k \in \mathbb{N}$.
    STEP 3 FOR $i = 0, 1, \ldots, I_k - 1$
        STEP 3.1 Choose $\alpha_{k,i} > 0$ and $D_{k,i} \in \mathcal{S}_{++}(\mathbb{R}^n)$.
        STEP 3.2 Compute $\tilde{y}^{(k,i)} \approx \hat{y}^{(k,i)} = \text{prox}_{\alpha_{k,i}\ell^{(k)}}^{D_{k,i}}(\tilde{x}^{(k,i)} - \alpha_{k,i} D_{k,i}^{-1} \nabla f_0(\tilde{x}^{(k,i)}))$
        such that
            $h^{(k,i)}(\tilde{y}^{(k,i)}) - h^{(k,i)}(\hat{y}^{(k,i)}) \leqslant -\frac{\tau}{2} h^{(k,i)}(\tilde{y}^{(k,i)})$.
        STEP 3.3 Set $d^{(k,i)} = \tilde{y}^{(k,i)} - \tilde{x}^{(k,i)}$.
        STEP 3.4 Compute $\lambda_{k,i} = \delta^{m_{k,i}}$, where $m_{k,i}$ is the smallest nonnegative integer such that
            $(f_0 + \ell^{(k)})(\tilde{x}^{(k,i)} + \delta^{m_{k,i}} d^{(k,i)}) \leqslant (f_0 + \ell^{(k)})(\tilde{x}^{(k,i)}) + \beta \delta^{m_{k,i}} h_\gamma^{(k,i)}(\tilde{y}^{(k,i)})$.
        STEP 3.5 Compute $\tilde{x}^{(k,i+1)} = \tilde{x}^{(k,i)} + \lambda_{k,i} d^{(k,i)}$.
    END
    STEP 4 Set

$$x^{(k+1)} = \begin{cases} \tilde{x}^{(k,I_k)} & \text{if } f(\tilde{x}^{(k,I_k)}) \leqslant f(\tilde{y}^{(k,0)}) \\ \tilde{y}^{(k,0)} & \text{otherwise} \end{cases}. \qquad (31)$$

END

3.3. Convergence analysis

We now develop the convergence analysis for algorithm 1. The main result, stated in theorem 1 to follow, establishes the convergence of the sequence $\{x^{(k)}\}_{k \in \mathbb{N}}$ to a stationary point of problem (21). Our convergence analysis can be considered as a generalization of the one devised for the VMILAn algorithm in [9], even though some major modifications are needed to establish a connection between the (possibly nonconvex) objective function and its iteratively reweighted convex majorizers.

In order to proceed with the analysis, we need some additional hypotheses. First, we state the assumptions needed on the algorithmic parameters $\alpha_{k,i}$, $D_{k,i}$, I_k, which are all required to belong to compact sets.

Assumption 2. *(i) There exists $I_{\max} \in \mathbb{Z}_+$ such that $0 < I_k \leqslant I_{\max}$ for all $k \geqslant 0$.*
(ii) There exist $0 < \alpha_{\min} \leqslant \alpha_{\max}$ such that $\alpha_{k,i} \in [\alpha_{\min}, \alpha_{\max}]$, for all $k \geqslant 0$, $j = 1, \ldots, J$.
(iii) There exist $\mu > 0$ such that $\frac{1}{\mu}\mathcal{I}_n \preceq D_{k,i} \preceq \mu\mathcal{I}_n$, for all $k \geqslant 0$, $j = 1, \ldots, J$.

Under the above assumptions, we can prove that the linesearch parameters $\lambda_{k,i}$ are uniformly bounded from below by a positive constant.

Lemma 7. *Suppose assumptions 1 and 2 hold. Then, there exists $\lambda_{\min} > 0$ such that*

$$\lambda_{k,i} \geqslant \lambda_{\min}, \quad \forall k \geqslant 0, \, j = 1, \ldots, J. \qquad (32)$$

Proof. The thesis follows by employing the same arguments as in [35, proposition 3.2]. $\square$

Next, we show that a sufficient decrease condition holds for the sequence of function values $\{f(x^{(k)})\}_{k \in \mathbb{N}}$ evaluated at the iterates of IR-VMILA. Furthermore, we show that the gap

between two consecutive function values $f(x^{(k)})$ and $f(x^{(k+1)})$ is a quantity converging to zero, and related to the distance between the two consecutive iterates $x^{(k)}$ and $x^{(k+1)}$.

Lemma 8. *Suppose assumptions 1 and 2 hold. Then the following statements hold.*

(i) For all $k \geqslant 0$, there holds:

$$f(x^{(k+1)}) + a \sum_{i=0}^{I_k-1} (-h^{(k,i)}(\tilde{y}^{(k,i)})) \leqslant f(x^{(k)}), \tag{33}$$

where $a = \beta\lambda_{\min} > 0$. Thus, the sequence $\{f(x^{(k)})\}_{k \in \mathbb{N}}$ is monotone nonincreasing.

(ii) We have:

$$\lim_{k \to \infty} \sum_{i=0}^{I_k-1} (-h^{(k,i)}(\tilde{y}^{(k,i)})) = 0. \tag{34}$$

(iii) For all $k \geqslant 0$, we have:

$$\|x^{(k+1)} - x^{(k)}\| \leqslant d \sqrt{\sum_{i=0}^{I_k-1} (-h^{(k,i)}(\tilde{y}^{(k,i)}))}, \tag{35}$$

where $d = 2\sqrt{I_{\max}\alpha_{\max}\mu(1+\tau)} > 0$.

Proof. (i) From STEPS 3.4 and 3.5 of algorithm 1, we have:

$$(f_0 + \ell^{(k)})(\tilde{x}^{(k,i+1)}) \leqslant (f_0 + \ell^{(k)})(\tilde{x}^{(k,i)}) + \beta\lambda_{k,i}h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}). \tag{36}$$

Summing (36) over $i = 0, 1, \ldots, I_k - 1$, simplifying terms, and recalling that $q^{(k)} = \ell^{(k)} + c^{(k)}$ (compare (24) with (25)) leads to:

$$(f_0 + q^{(k)})(\tilde{x}^{(k,I_k)}) \leqslant (f_0 + q^{(k)})(\tilde{x}^{(k,0)}) + \beta \sum_{i=0}^{I_k-1} \lambda_{k,i}h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}). \tag{37}$$

Since $q^{(k)}$ is a majorant of f_1 at point $x^{(k)}$, and due to STEPS 2 and 4 of algorithm 1, we can write the following inequalities:

$$(f_0 + q^{(k)})(\tilde{x}^{(k,I_k)}) \geqslant f(\tilde{x}^{(k,I_k)}) \geqslant f(x^{(k+1)}) \tag{38}$$

$$(f_0 + q^{(k)})(\tilde{x}^{(k,0)}) = f(\tilde{x}^{(k,0)}) = f(x^{(k)}). \tag{39}$$

By applying (38) and (39) to (37), we come to:

$$f(x^{(k+1)}) \leqslant f(x^{(k)}) + \beta \sum_{i=0}^{I_k-1} \lambda_{k,i}h_\gamma^{(k,i)}(\tilde{y}^{(k,i)}). \tag{40}$$

From inequality (40), (29), $\lambda_{k,i} \geqslant \lambda_{\min}$ (see lemma 7), and (12), we get (33). Furthermore, it follows from (40) and (29) that the sequence $\{f(x^{(k)})\}_{k \in \mathbb{N}}$ is monotone nondecreasing.

(ii) Due to assumption 1, f is continuous and coercive, hence it is bounded from below. This implies that the monotone nondecreasing sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ is bounded from below. Then, there must exist $\underline{f}\in\mathbb{R}$ such that $f(x^{(k)})\to\underline{f}$, and taking the limit of (33) over $k\to\infty$ we obtain (34).

(iii) If STEP 4 yields $x^{(k+1)}=\tilde{x}^{(k,I_k)}$, then an easy application of Jensen's inequality shows that:

$$\|x^{(k+1)}-x^{(k)}\|^2 = \left\|\sum_{i=0}^{I_k-1}(\tilde{x}^{(k,i+1)}-\tilde{x}^{(k,i)})\right\|^2 \leqslant I_k\sum_{i=0}^{I_k-1}\|\tilde{x}^{(k,i+1)}-\tilde{x}^{(k,i)}\|^2$$

$$\leqslant I_{\max}\sum_{i=0}^{I_k-1}\|\tilde{x}^{(k,i)}-\tilde{y}^{(k,i)}\|^2,$$

where the last inequality is due to STEP 3.5 of algorithm 1, $\lambda_{k,i}\leqslant 1$, and assumption 2(i). Otherwise, if $x^{(k+1)}=\tilde{y}^{(k,0)}$, then it trivially follows that:

$$\|x^{(k+1)}-x^{(k)}\|^2 = \|\tilde{y}^{(k,0)}-\tilde{x}^{(k,0)}\|^2 \leqslant I_{\max}\sum_{i=0}^{I_k-1}\|\tilde{x}^{(k,i)}-\tilde{y}^{(k,i)}\|^2. \tag{41}$$

Thus, in both cases, inequality (41) holds. Finally, applying inequality (17) in lemma 5 with $h=h^{(k,i)}$, $\tilde{y}=\tilde{y}^{(k,i)}$, $x=\tilde{x}^{(k,i)}$ yields the thesis.

$\square$

In the following lemma, we prove that the norm of the subdifferential of f at the exact proximal–gradient point $\hat{y}^{(k,0)}$ can be suitably upper bounded by a quantity converging to zero. In the KL framework, this is usually referred to as the relative error condition [8]. Unlike other standard results available in the literature, our result holds at the exact proximal–gradient point $\hat{y}^{(k,0)}$, of which we have no knowledge if $\tau>0$, rather than at the point $\tilde{y}^{(k,0)}$ that is actually computed in the algorithm. This is because such an upper bound is unlikely to hold at an inexact point computed according to (28), as discussed in [36, section 3.2]. Indeed, even if it is possible to obtain an upper bound for the distance between the exact point $\hat{y}^{(k,0)}$ and its approximation $\tilde{y}^{(k,0)}$ (see inequality (16) in lemma 5), a similar relation between the subgradients at $\hat{y}^{(k,0)}$ and those at $\tilde{y}^{(k,0)}$ can not be proved.

Lemma 9. *Suppose assumptions 1 and 2 hold. For all $k\geqslant 0$, there exists $v^{(k)}\in\partial f(\hat{y}^{(k,0)})$ such that:*

$$\|v^{(k)}\| \leqslant \tilde{b}\sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))}, \tag{42}$$

where $\tilde{b}=\left(L+\frac{\mu}{\alpha_{\min}}+\rho_K\xi_K J\right)\sqrt{2\alpha_{\max}\mu\left(1+\frac{\tau}{2}\right)}>0$.

Proof. By the definition (27) of exact proximal–gradient point, the proximal point $\hat{y}^{(k,0)}$ is defined as:

$$\hat{y}^{(k,0)}=\underset{z\in\mathbb{R}^n}{\operatorname{argmin}}\,h^{(k,i)}(z) \quad\Leftrightarrow\quad 0\in\partial h^{(k,i)}(\hat{y}^{(k,0)}).$$

By applying lemma 2(i) to the differential inclusion $0 \in \partial h^{(k,i)}(\hat{y}^{(k,0)})$, we obtain that the following vector $u^{(k)} \in \mathbb{R}^n$ is a subgradient of $\ell^{(k)}$ at point $\hat{y}^{(k,0)}$:

$$u^{(k)} := \frac{1}{\alpha_{k,0}} D_{k,0}(\tilde{x}^{(k,0)} - \alpha_{k,0} D_{k,0}^{-1} \nabla f_0(\tilde{x}^{(k,0)}) - \hat{y}^{(k,0)}) \in \partial \ell^{(k)}(\hat{y}^{(k,0)}). \qquad (43)$$

Recalling the definition of $\ell^{(k)}$ in (25) and using lemma 2(ii) and (iii), it follows that:

$$\partial \ell^{(k)}(\hat{y}^{(k,0)}) = \partial \left(\sum_{j=1}^{J} (\phi'_j \circ \psi_j)(x^{(k)}) \psi_j + \chi \right)(\hat{y}^{(k,0)})$$

$$= \sum_{j=1}^{J} (\phi'_j \circ \psi_j)(x^{(k)}) \partial \psi_j(\hat{y}^{(k,0)}) + \partial \chi(\hat{y}^{(k,0)}),$$

which implies the existence of J vectors $u_j(\hat{y}^{(k,0)}) \in \partial \psi_j(\hat{y}^{(k,0)}), j = 1,\ldots,J$, and $u_\chi(\hat{y}^{(k,0)}) \in \partial \chi(\hat{y}^{(k,0)})$ such that:

$$u^{(k)} := \sum_{j=1}^{J} (\phi'_j \circ \psi_j)(x^{(k)}) \cdot u_j(\hat{y}^{(k,0)}) + u_\chi(\hat{y}^{(k,0)}). \qquad (44)$$

Using lemma 2(i), we can write $\partial f(\hat{y}^{(k,0)}) = \nabla f_0(\hat{y}^{(k,0)}) + \partial f_1(\hat{y}^{(k,0)})$. Note that f_1 can be written as $f_1 = \phi \circ \Psi + \chi$, where $\Psi(x) = (\psi_1(x),\ldots,\psi_J(x))$ and $\phi(z) = \sum_{j=1}^{J} \phi(z_j)$ satisfy the hypotheses of lemma 3 thanks to assumption 1(ii). Therefore, we can apply lemma 3 to obtain,

$$\partial f_1(\hat{y}^{(k,0)}) = \sum_{j=1}^{J} (\phi_j \circ \psi_j)(\hat{y}^{(k,0)}) \partial \psi_j(\hat{y}^{(k,0)}) + \partial \chi(\hat{y}^{(k,0)}).$$

Hence, the following vector $w^{(k)} \in \mathbb{R}^n$ is a subgradient of f_1 at point $\hat{y}^{(k,0)}$:

$$w^{(k)} = \sum_{j=1}^{J} (\phi'_j \circ \psi_j)(\hat{y}^{(k,0)}) \cdot u_j(\hat{y}^{(k,0)}) + u_\chi(\hat{y}^{(k,0)}) \in \partial f_1(\hat{y}^{(k,0)}), \qquad (45)$$

and $v^{(k)} := \nabla f_0(\hat{y}^{(k,0)}) + w^{(k)}$ is a subgradient of f at point $\hat{y}^{(k,0)}$, i.e. $v^{(k)} \in \partial f(\hat{y}^{(k,0)})$.

By lemma 8, it follows that the sequence $\{x^{(k)}\}_{k \in \mathbb{N}}$ is contained in the set $K_1 = [f \leqslant f(x^{(0)})]$, which is compact due to assumption 1(iii). Analogously, denoting $\omega_{\max} = \max_{k \in \mathbb{N}} |\phi' \circ \psi_j(x^{(k)})|$, which is a finite real value due to the continuity of $\phi' \circ \psi_j$ on K_1, we note that $\{\hat{y}^{(k,0)}\}_{k \in \mathbb{N}}$ is a subset of the set:

$$K_2 = \{ \mathrm{prox}_{\sum_{j=1}^{J} \alpha \omega_j \psi_j + \chi}^{D}(x - \alpha D^{-1} \nabla f_0(x)) : \ x \in K_1, \alpha \in [\alpha_{\min}, \alpha_{\max}],$$

$$\frac{1}{\mu} I \preceq D \preceq \mu I, \ |\omega_j| \leqslant \omega_{\max} \}, \qquad (46)$$

which is compact due to the continuity of the proximal operator with respect to its arguments, and the compactness of K_1. Thus, letting $K := K_1 \cup K_2$, we have that the inclusion $\{x^{(k)}\}_{k \in \mathbb{N}} \cup \{\hat{y}^{(k,0)}\}_{k \in \mathbb{N}} \subseteq K$ holds.

Now, we can write the following inequalities:

$$
\begin{aligned}
\|v^{(k)}\| &= \|\nabla f_0(\hat{y}^{(k,0)}) + w^{(k)}\| \\
&\leqslant \|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| + \|w^{(k)} - u^{(k)}\| \\
&= \|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| + \left\|\sum_{j=1}^{J}((\phi_j' \circ \psi_j)(\hat{y}^{(k,0)}) - (\phi_j' \circ \psi_j)(x^{(k)}))u_j(\hat{y}^{(k,0)})\right\| \\
&\leqslant \|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| + \sum_{j=1}^{J}|(\phi_j' \circ \psi_j)(\hat{y}^{(k,0)}) - (\phi_j' \circ \psi_j)(x^{(k)})|\|u_j(\hat{y}^{(k,0)})\| \\
&\leqslant \|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| + \rho_K \xi_K J \|\hat{y}^{(k,0)} - x^{(k)}\| \\
&\leqslant \|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| + \rho_K \xi_K J \sqrt{2\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)}\sqrt{-h^{(k,0)}(\tilde{y}^{(k,0)})},
\end{aligned}
\tag{47}
$$

where the first equality is due to the definition of $v^{(k)}$, the first inequality follows from the triangular inequality, the second equality comes from (44) and (45), the second inequality is again due to the triangular inequality, whereas the third one follows from remark 4(ii) and (iii), and finally the fourth one is a consequence of inequality (15) in lemma 5 applied with $h = h^{(k,0)}, \hat{y} = \hat{y}^{(k,0)}, \tilde{y} = \tilde{y}^{(k,0)} x = x^{(k)}$.

Finally, we focus on the quantity $\|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\|$, which can be upper bounded in the following way:

$$
\begin{aligned}
\|\nabla f_0(\hat{y}^{(k,0)}) + u^{(k)}\| &= \left\|\nabla f_0(\hat{y}^{(k,0)}) + \frac{1}{\alpha_{k,0}}D_{k,0}(\tilde{x}^{(k,0)} - \alpha_{k,0}D_{k,0}^{-1}\nabla f_0(\tilde{x}^{(k,0)}) - \hat{y}^{(k,0)})\right\| \\
&= \left\|\nabla f_0(\hat{y}^{(k,0)}) - \nabla f_0(\tilde{x}^{(k,0)}) + \frac{1}{\alpha_{k,0}}D_{k,0}(\tilde{x}^{(k,0)} - \hat{y}^{(k,0)})\right\| \\
&\leqslant \left(L + \frac{\mu}{\alpha_{\min}}\right)\|\hat{y}^{(k,0)} - \tilde{x}^{(k,0)}\| \\
&\leqslant \left(L + \frac{\mu}{\alpha_{\min}}\right)\sqrt{2\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)}\sqrt{-h^{(k,0)}(\tilde{y}^{(k,0)})},
\end{aligned}
\tag{48}
$$

where the first equality follows from (43), the first inequality from the triangular inequality, the Lipschitz continuity of ∇f_0 (see assumption 1(i)), and assumptions 2(ii) and (iii), whereas the last inequality is due again to inequality (15) in lemma 5 applied with $h = h^{(k,0)}, \hat{y} = \hat{y}^{(k,0)}, \tilde{y} = \tilde{y}^{(k,0)} x = x^{(k)}$. Combining (47) and (48) easily yields the thesis. $\qquad\square$

The next lemma contains the first convergence result for IR-VMILA, i.e. that each limit point of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is stationary for problem (21). Even though the arguments employed are analogous to those followed in several other algorithms framed within the KL framework [3, 4, 8, 27], we report its proof for the sake of completeness.

Lemma 10. *Suppose assumptions 1 and 2 hold true. Let $X^*(x^{(0)})$ denote the set of all limit points of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ generated by algorithm 1 starting from the initial iterate $x^{(0)}$. Then, the following statements hold true.*

(i) $X^*(x^{(0)})$ is nonempty and compact.
(ii) There holds $\lim_{k\to\infty}\mathrm{dist}(x^{(k)}, X^*(x^{(0)})) = 0$.
(iii) We have $\lim_{k\to\infty}f(x^{(k)}) = f(x^*)$.

(iv) There exists $f^* \in \mathbb{R}$ such that $f(x^*) = f^*$ for all $x^* \in X^*(x^{(0)})$.

(v) If $x^* \in X^*(x^{(0)})$, then x^* is a stationary point for problem (21).

Proof. (i) From lemma 8(i), we know that $\{x^{(k)}\}_{k\in\mathbb{N}} \subseteq [f \leqslant f(x^{(0)})]$. Then, assumption 1(iii) implies that $\{x^{(k)}\}_{k\in\mathbb{N}}$ is bounded, which in turn guarantees the existence of at least one limit point, hence $X^*(x^{(0)}) \neq \emptyset$. The compactness of $X^*(x^{(0)})$ holds by observing that $X^*(x^{(0)})$ is a countable intersection of compact sets, exactly as in [20, lemma 5].

(ii) This is a consequence of the definition of limit point of $\{x^{(k)}\}_{k\in\mathbb{N}}$.

(iii) Let $x^* \in X^*(x^{(0)})$. Since $x^{(k)} \in \mathrm{dom}(\chi)$ for all $k \geqslant 0$, and f is continuous on $\mathrm{dom}(\chi)$ by assumption 1, it follows that there exists a subset of indices $K \subseteq \mathbb{N}$ such that $\lim_{k\in K} f(x^{(k)}) = f(x^*)$. As $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ is a monotone nonincreasing and bounded sequence, it converges to a finite value, hence we have $\lim_{k\to\infty} f(x^{(k)}) = f(x^*)$.

(iv) This is a straightforward consequence of point (iii).

(v) Let $x^* \in X^*(x^{(0)})$ and $K \subseteq \mathbb{N}$ any subset of indices such that $\lim_{k\in K} x^{(k)} = x^*$. From point (iii), we also know that $\lim_{k\in K} f(x^{(k)}) = f(x^*)$. In addition, by plugging lemma 8(ii) inside (42), we obtain that $\lim_{k\in K} v^{(k)} = 0$. Hence, we can employ the closedness property of the limiting subdifferential (see lemma 1) to conclude that $0 \in \partial f(x^*)$. $\square$

As in [9], we now introduce the following surrogate function:

$$F : \mathbb{R}^n \times \mathbb{R} \to \bar{\mathbb{R}}$$
$$F(z,\sigma) = f(z) + \frac{1}{2}\sigma^2, \tag{49}$$

which will be needed to incorporate the inexact computation of point $\tilde{y}^{(k,0)}$ inside the proof of theorem 1.

In what follows, we prove that a relative error condition holds for the subdifferential of F evaluated at the point $(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)})$, where $\{\tilde{\sigma}^{(k,0)}\}_{k\in\mathbb{N}}$ is a suitable auxiliary sequence converging to zero. Furthermore, we show that the sequence $\{F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)})\}_{k\in\mathbb{N}}$ is controlled from below and above by the function values $\{f(x^{(k)})\}_{k\in\mathbb{N}}$.

Lemma 11. *Suppose assumptions 1 and 2 hold. Let $c \geqslant 0$ and $\tilde{b} > 0$ be the constants defined in lemmas 6 and 9, respectively. For all $k \geqslant 0$, define:*

$$\tilde{\sigma}^{(k,0)} = \sqrt{2(c(-h^{(k,0)}(\tilde{y}^{(k,0)})) + q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}))}. \tag{50}$$

Furthermore, let $b = \tilde{b} + \sqrt{2c} + \sqrt{4\rho_K \xi_K \alpha_{\max} \mu \left(1 + \frac{\tau}{2}\right)}$, where $\tilde{b} > 0$ is the constant defined in lemma 9.

(i) We have $\lim_{k\to\infty} \tilde{\sigma}^{(k,0)} = 0$.

(ii) For all $k \geqslant 0$, there holds:

$$\mathrm{dist}(0, \partial F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)})) \leqslant b\sqrt{\sum_{i=0}^{I_k-1} (-h^{(k,i)}(\tilde{y}^{(k,i)}))}. \tag{51}$$

(iii) There exists a sequence $\{r^{(k)}\}_{k\in\mathbb{N}} \subseteq \mathbb{R}$ such that $\lim_{k\to\infty} r^{(k)} = 0$ and, for all $k \geqslant 0$, there holds:

$$f(x^{(k+1)}) \leqslant F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) \leqslant f(x^{(k)}) + r^{(k)}. \tag{52}$$

Proof. (i) Let $K_1 = [f \leqslant f(x^{(0)})]$, $K_2 \subseteq \mathbb{R}^n$ defined as in (46), and $K = K_1 \cup K_2$. Let $\xi_K > 0$ and ρ_K be the local Lipschitz constants common to all functions ψ_j and $\phi_j' \circ \psi_j$,

$j = 1, \ldots, J$, respectively, on the compact set K (see remark 4). Then, by the definition of $q^{(k)}$ in (24), the concavity of functions ϕ_j, the local Lipschitz continuity of functions ϕ_j and $\phi_j' \circ \psi_j$, and inequality (15) with $\hat{y} = \hat{y}^{(k,0)}$, $\tilde{y} = \tilde{y}^{(k,0)}$, and $x = x^{(k)}$, we can deduce the following chain of inequalities:

$$
\begin{aligned}
0 \leqslant q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}) &= \sum_{j=1}^{J}((\phi_j \circ \psi_j)(x^{(k)}) - (\phi_j \circ \psi_j)(\hat{y}^{(k,0)})) \\
&\quad + (\phi_j' \circ \psi_j)(x^{(k)})(\psi_j(\hat{y}^{(k,0)}) - \psi_j(x^{(k)})) \\
&\leqslant \sum_{j=1}^{J}(\phi_j' \circ \psi_j)(\hat{y}^{(k,0)})(\psi_j(x^{(k)}) - \psi_j(\hat{y}^{(k,0)})) \\
&\quad + (\phi_j' \circ \psi_j)(x^{(k)})(\psi_j(\hat{y}^{(k,0)}) - \psi_j(x^{(k)})) \\
&= \sum_{j=1}^{J}(\phi_j' \circ \psi_j(\hat{y}^{(k,0)}) - \phi_j' \circ \psi_j(x^{(k)}))(\psi_j(x^{(k)}) - \psi_j(\hat{y}^{(k,0)})) \\
&\leqslant \rho_K \xi_K \|\hat{y}^{(k,0)} - x^{(k)}\|^2 \\
&\leqslant 2\rho_K \xi_K \alpha_{\max} \mu \left(1 + \frac{\tau}{2}\right)(-h^{(k,0)}(\tilde{y}^{(k,0)})).
\end{aligned}
\tag{53}
$$

By employing lemma 8(ii) in the previous inequality, we obtain that $\lim_{k \to \infty} q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}) = 0$, and using the definition of $\tilde{\sigma}^{(k,0)}$ in (50), the thesis follows.

(ii) Let $v^{(k)} \in \partial f(\hat{y}^{(k,0)})$ be the subgradient in lemma 9. Then, it is easily follows that:

$$
\begin{pmatrix} v^{(k)} \\ \tilde{\sigma}^{(k,0)} \end{pmatrix} \in \partial F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}).
$$

Using the triangular inequality, lemma 9, and the definition of $\tilde{\sigma}^{(k,0)}$ in (50), we obtain:

$$
\begin{aligned}
\left\| \begin{pmatrix} v^{(k)} \\ \tilde{\sigma}^{(k,0)} \end{pmatrix} \right\| &\leqslant \|v^{(k)}\| + |\tilde{\sigma}^{(k,0)}| \\
&\leqslant \tilde{b}\sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} + \sqrt{2(c(-h^{(k,0)}(\tilde{y}^{(k,0)})) + q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}))} \\
&\leqslant \tilde{b}\sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} + \sqrt{2c(-h^{(k,0)}(\tilde{y}^{(k,0)}))} \\
&\quad + \sqrt{2(q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}))}.
\end{aligned}
$$

The thesis now follows by applying (53) to the previous inequality.

(iii) We start by observing that:

$$
\begin{aligned}
F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) &= f(\hat{y}^{(k,0)}) + c(-h^{(k,0)}(\tilde{y}^{(k,0)})) + q^{(k)}(\hat{y}^{(k,0)}) - f_1(\hat{y}^{(k,0)}) \\
&= f_0(\hat{y}^{(k,0)}) + q^{(k)}(\hat{y}^{(k,0)}) + c(-h^{(k,0)}(\tilde{y}^{(k,0)})).
\end{aligned}
\tag{54}
$$

First, we prove the left inequality in (52). From (18) in lemma 6 with $\ell = \ell^{(k)}$, $h(\cdot;x) = h^{(k,0)}(\cdot)$, $\hat{y} = \hat{y}^{(k,0)}$, $\tilde{y} = \tilde{y}^{(k,0)}$, $x = x^{(k)}$, we can write:

$$(f_0 + \ell^{(k)})(\hat{y}^{(k,0)}) \geqslant (f_0 + \ell^{(k)})(\tilde{y}^{(k,0)}) - c(-h^{(k,0)}(\tilde{y}^{(k,0)})),$$

or equivalently (see the definition of $q^{(k)}$ in (24)):

$$(f_0 + q^{(k)})(\hat{y}^{(k,0)}) \geqslant (f_0 + q^{(k)})(\tilde{y}^{(k,0)}) - c(-h^{(k,0)}(\tilde{y}^{(k,0)})). \tag{55}$$

Then we have:

$$F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) \geqslant (f_0 + q^{(k)})(\tilde{y}^{(k,0)}) \geqslant f(\tilde{y}^{(k,0)}) \geqslant f(x^{(k+1)}),$$

where the first inequality follows by plugging (55) inside (54), the second one follows from the fact that $q^{(k)}$ is a majorant for f_1 at $x^{(k)}$, whereas the third inequality is due to STEP 4 of algorithm 1.

As for the right inequality in (52), we apply (20) with $\ell = \ell^{(k)}$, $h(\cdot;x) = h^{(k,0)}(\cdot)$, $\hat{y} = \hat{y}^{(k,0)}$, $\tilde{y} = \tilde{y}^{(k,0)}$, $x = x^{(k)}$, to obtain:

$$(f_0 + \ell^{(k)})(\hat{y}^{(k,0)}) \leqslant (f_0 + \ell^{(k)})(x^{(k)}) - L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)h^{(k,0)}(\tilde{y}^{(k,0)}),$$

or equivalently

$$(f_0 + q^{(k)})(\hat{y}^{(k,0)}) \leqslant (f_0 + q^{(k)})(x^{(k)}) - L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)h^{(k,0)}(\tilde{y}^{(k,0)})). \tag{56}$$

Plugging (56) inside (54) and recalling that $q^{(k)}$ is a majorant for f_1 at $x^{(k)}$, we get:

$$F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) \leqslant (f_0 + q^{(k)})(x^{(k)}) + \left(c + L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)\right)(-h^{(k,0)}(\tilde{y}^{(k,0)}))$$

$$= f(x^{(k)}) + \left(c + L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)\right)(-h^{(k,0)}(\tilde{y}^{(k,0)})).$$

Setting $r^{(k)} = \left(c + L\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)\right)(-h^{(k,0)}(\tilde{y}^{(k,0)}))$ and recalling lemma 8(ii), we come to the second inequality in (52). $\square$

We are now ready to state the main result, namely that the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ satisfies the finite length property (which implies convergence to a stationary point). To this aim, we need the following additional requirement on F.

Assumption 3. *The surrogate function F defined in (49) is a KL function.*

The previous assumption is satisfied if f is the sum of semialgebraic, subanalytic and real analytic functions, see [9, remark 5].

Theorem 1. *Suppose that assumptions 1–3 hold. The sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ generated by algorithm 1 has finite length, i.e.*

$$\sum_{k=0}^{\infty}\|x^{(k+1)} - x^{(k)}\| < \infty,$$

hence it converges to a stationary point x^* of problem (21).

Proof. Suppose that there exists $\bar{k} \geqslant 0$ for which $\sum_{i=0}^{I_{\bar{k}}-1}(-h^{(\bar{k},i)}(\tilde{y}^{(\bar{k},i)})) = 0$. Then, inequality (42) implies that $0 \in \partial f(\hat{y}^{(\bar{k},0)})$, i.e. $\hat{y}^{(\bar{k},0)}$ is a stationary point for problem (21). In addition, from inequalities (15)–(17) with $h = h^{(\bar{k},0)}$, $\tilde{y} = \tilde{y}^{(\bar{k},0)}$, $x = x^{(\bar{k})}$, it must be $\hat{y}^{(\bar{k},0)} = \tilde{y}^{(\bar{k},0)} = x^{(\bar{k})}$, which in turn implies $d^{(\bar{k},0)} = 0$ and thus $x^{(\bar{k},1)} = x^{(\bar{k})}$. By induction, it follows that $x^{(\bar{k}+k)} = x^{(\bar{k})}$ for all $k \geqslant 0$, hence the sequence converges to the stationary point $x^{(\bar{k})}$.

Otherwise, let us assume that:

$$\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)})) > 0, \quad \forall\, k \geqslant 0, \tag{57}$$

and define the set:

$$\Omega^*(x^{(0)}) = X^*(x^{(0)}) \times \{0\},$$

where $X^*(x^{(0)})$ is the set of all limit points of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$. Then, for all $v^* \in \Omega^*(x^{(0)})$, where $v^* = (x^*,0)$ with $x^* \in X^*(x^{(0)})$, we have:

$$F(v^*) = f(x^*) = f^*, \quad \forall\, v^* \in \Omega^*(x^{(0)}),$$

thanks to the definition of the surrogate function F in (49) and lemma 10(iv).

Therefore, since F is a KL function (assumption 3), and F is constant and equal to f^* over the set of limit points $\Omega^*(x^{(0)})$, which is compact (lemma 10(i)), there exist $\rho, \nu > 0$ and $\xi : [0, \nu) \to [0, +\infty)$ as in lemma 4 such that (9) holds for all $x \in B^*$, where B^* is defined in (10) with $X = \Omega^*(x^{(0)})$.

Now, it is easy to show that the pair $(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)})$ belongs to B^* for sufficiently large k. Indeed, we have:

$$\operatorname{dist}(\hat{y}^{(k,0)}, X^*(x^{(0)})) \leqslant \|\hat{y}^{(k,0)} - x^{(k)}\| + \operatorname{dist}(x^{(k)}, X^*(x^{(0)}))$$
$$\leqslant \sqrt{2\alpha_{\max}\mu\left(1 + \frac{\tau}{2}\right)(-h^{(k,0)}(\tilde{y}^{(k,0)}))} + \operatorname{dist}(x^{(k)}, X^*(x^{(0)})),$$

where the second inequality is a consequence of (15) in lemma 5 applied with $h = h^{(k,0)}$, $\hat{y} = \hat{y}^{(k,0)}$, $\tilde{y} = \tilde{y}^{(k,0)}$ $x = x^{(k)}$. By definition of the set $X^*(x^{(0)})$ and lemma 8(ii), the previous inequality implies the existence of an index $k_1 \geqslant 0$ such that:

$$\operatorname{dist}(\hat{y}^{(k,0)}, X^*(x^{(0)})) < \frac{\rho}{2}, \quad \forall\, k \geqslant k_1. \tag{58}$$

By lemma 11(i), there exists also $k_2 \geqslant 0$ such that:

$$\tilde{\sigma}^{(k,0)} < \frac{\rho}{2}, \quad \forall\, k \geqslant k_2. \tag{59}$$

Furthermore, combining lemmas 10(iii) and 11(iii) yields an index $k_3 \geqslant 0$ such that:

$$f^* < f(x^{(k+1)}) \leqslant F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) < f^* + \nu, \quad \forall\, k \geqslant k_3, \tag{60}$$

where the strict inequality on the left is motivated by the fact that (33) and (57) imply $f(x^{(k+1)}) \geqslant f(x^{(k+2)}) \geqslant f^*$. Thus, combining inequalities (58)–(60) and setting $k_0 = \max\{k_1, k_2, k_3\}$, we conclude that:

$$\operatorname{dist}((\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}), \Omega^*(x^{(0)})) \leqslant \operatorname{dist}(\hat{y}^{(k,0)}, X^*(x^{(0)})) + |\tilde{\sigma}^{(k,0)}| < \rho$$
$$f^* < F(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) < f^* + \nu,$$

for all $k \geqslant k_0$. If we denote by $\{x^{(k)}\}_{k\in\mathbb{N}}$ the sequence $\{x^{(k+k_0)}\}_{k\in\mathbb{N}}$, then the previous inequalities are equivalent to the following inclusion:

$$(\hat{y}^{(k,0)}, \tilde{\sigma}^{(k,0)}) \in B^*, \quad \forall\, k \geqslant 0.$$

From the previous inclusion and lemma 4, we can apply the uniformized KL inequality (9) with $x = (\hat{y}^{(k-1,0)}, \tilde{\sigma}^{(k-1,0)})$ for all $k \geqslant 1$, i.e.

$$\xi'(F(\hat{y}^{(k-1,0)}, \tilde{\sigma}^{(k-1,0)}) - f^*)\operatorname{dist}(0, \partial F(\hat{y}^{(k-1,0)}, \tilde{\sigma}^{(k-1,0)})) \geqslant 1, \quad \forall\, k \geqslant 1.$$

Since ξ is concave and differentiable, ξ' is nonincreasing. Then, inequality (60) implies:

$$\xi'(f(x^{(k)}) - f^*) \geqslant \xi'(F(\hat{y}^{(k-1,0)}, \tilde{\sigma}^{(k-1,0)}) - f^*),$$

and hence:

$$\xi'(f(x^{(k)}) - f^*)\text{dist}(0, \partial F(\hat{y}^{(k-1,0)}, \tilde{\sigma}^{(k-1,0)})) \geqslant 1, \quad \forall\, k \geqslant 1. \tag{61}$$

Combining (61) and lemma 11(ii) leads to:

$$\xi'(f(x^{(k)}) - f^*) \geqslant \frac{1}{\|v^{(k-1)}\|} \geqslant \frac{1}{b\sqrt{\sum_{i=0}^{I_{k-1}-1}\left(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)})\right)}}. \tag{62}$$

Next, we can write the following chain of inequalities:

$$\begin{aligned}
\xi(f(x^{(k)}) - f^*) - \xi(f(x^{(k+1)}) - f^*) &\geqslant \xi'(f(x^{(k)}) - f^*)(f(x^{(k)}) - f(x^{(k+1)})) \\
&\geqslant \frac{f(x^{(k)}) - f(x^{(k+1)})}{b\sqrt{\sum_{i=0}^{I_{k-1}-1}\left(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)})\right)}} \\
&\geqslant \frac{a\sum_{i=0}^{I_k-1}\left(-h^{(k,i)}(\tilde{y}^{(k,i)})\right)}{b\sqrt{\sum_{i=0}^{I_{k-1}-1}\left(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)})\right)}},
\end{aligned} \tag{63}$$

where the first inequality is due to the concavity of ξ, the second one follows from (62), and the third one from (33). Setting:

$$s^{(k)} = \frac{b}{a}(\xi(f(x^{(k)}) - f^*) - \xi(f(x^{(k+1)}) - f^*)),$$

the previous inequality can be reformulated as:

$$\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)})) \leqslant s^{(k)}\sqrt{\sum_{i=0}^{I_{k-1}-1}\left(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)})\right)}.$$

By applying the square root on both sides and using the basic inequality $2\sqrt{uv} \leqslant u + v$ with $u = s^{(k)}$ and $v = \sqrt{\sum_{i=0}^{I_{k-1}-1}(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)}))}$, we obtain:

$$2\sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} \leqslant \sqrt{\sum_{i=0}^{I_{k-1}-1}\left(-h^{(k-1,i)}(\tilde{y}^{(k-1,i)})\right)} + s^{(k)}.$$

Summing the previous inequality over $p = 1, \dots, k$ yields:

$$2\sum_{p=1}^{k}\sqrt{\sum_{i=0}^{I_p-1}(-h^{(p,i)}(\tilde{y}^{(p,i)}))} \leqslant \sum_{p=1}^{k}\sqrt{\sum_{i=0}^{I_{p-1}-1}\left(-h^{(p-1,i)}(\tilde{y}^{(p-1,i)})\right)} + \sum_{p=1}^{k}s^{(p)},$$

which by simplifying terms reduces to:

$$\sum_{p=1}^{k}\sqrt{\sum_{i=0}^{I_p-1}(-h^{(p,i)}(\tilde{y}^{(p,i)}))} + \sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} \leqslant \sqrt{\sum_{i=0}^{I_0-1}(-h^{(0,i)}(\tilde{y}^{(0,i)}))} + \sum_{p=1}^{k}s^{(p)}. \tag{64}$$

Note that:

$$\sum_{p=1}^{k} s^{(p)} = \frac{b}{a}(\xi(f(x^{(1)}) - f^*) - \xi(f(x^{(k+1)}) - f^*)) \leqslant \frac{b}{a}\xi(f(x^{(1)}) - f^*), \qquad (65)$$

where the second inequality follows from the fact that ξ has positive sign. Thus, combining (64) and (65) yields:

$$\sum_{p=1}^{k} \sqrt{\sum_{i=0}^{I_p-1}(-h^{(p,i)}(\tilde{y}^{(p,i)}))} \leqslant \sqrt{\sum_{i=0}^{I_0-1}(-h^{(0,i)}(\tilde{y}^{(0,i)}))} + \frac{b}{a}\xi(f(x^{(1)}) - f^*).$$

Taking the limit for $k \to \infty$ allows to conclude that:

$$\sum_{k=0}^{\infty} \sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} < \infty.$$

Finally, recalling lemma 8(iii), we can conclude that:

$$\sum_{k=0}^{\infty} \|x^{(k+1)} - x^{(k)}\| \leqslant d\sum_{k=0}^{\infty} \sqrt{\sum_{i=0}^{I_k-1}(-h^{(k,i)}(\tilde{y}^{(k,i)}))} < \infty. \qquad (66)$$

By (66) and lemma 10(v), the thesis follows. $\qquad \square$

The following result is concerned with the convergence rate of the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$, under the hypothesis that the desingularizing function ξ is defined as $\xi(t) = \frac{1}{c}t^{1-\theta}$, with $c > 0$ and $\theta \in (0,1)$. In particular, it is stated that the iterates $\{x^{(k)}\}_{k\in\mathbb{N}}$ converge either polynomially, linearly, or in a finite number of iterations, depending on the value that the KL exponent θ assumes.

Theorem 2. *Suppose that assumptions 1–3 hold. Let x^* be the unique limit point of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ generated by algorithm 1, and suppose that assumption 3 is satisfied with $\xi(t) = \frac{1}{c}t^{1-\theta}$, being $c > 0$ and $\theta \in (0,1)$.*

(i) If $\theta \in (\frac{1}{2},1)$, then there exists $c_1 > 0$ such that:
$$f(x^{(k)}) - f(x^*) \leqslant c_1 \left(\frac{1}{k}\right)^{\frac{1}{2\theta-1}},$$
for all sufficiently large k.

(ii) If $\theta \in (0,\frac{1}{2}]$, then there exists $c_2 > 0$ such that:
$$f(x^{(k)}) - f(x^*) \leqslant c_2 \left(\frac{1}{e}\right)^{k},$$
for all sufficiently large k.

(iii) If $\theta = 0$, then $\{x^{(k)}\}_{k\in\mathbb{N}}$ converges in a finite number of iterations.

Proof. The proof is omitted, as it runs analogously to the one available for the VMILAn algorithm in [9, theorem 2] by combining lemmas 8 and 11, the KL property, and the monotonicity of the derivative ξ'. $\qquad \square$

4. Numerical experiments

In order to evaluate the performance of IR-VMILA, we turn to the realm of imaging inverse problems, as many of them can be expressed as optimization problems. In this context, the use of composite energy functionals with one or more nonconvex terms, that better capture the properties of the images of interest, has become extremely popular in the last few years [13, 16, 23, 27].

The following experiments are implemented with Matlab 2022a on a Macbook Pro equipped with a 2,4 GHz Intel Core i5 quad-core CPU.

4.1. Image deblurring and denoising using smoothed TV and bimodal regularization

In image deconvolution the goal is to compute a suitable approximation $\bar{x}$ of the true unobtainable ground truth $x^{true} \in \mathbb{R}^n$, which gets corrupted with blur and noise in the acquisition process. The observed data can then be modelled as $y = Hx^{true} + n$, where $H \in \mathbb{R}^{n \times n}$ is the matrix representing the convolution with the filter h causing the blurring effect, i.e. the point spread function (PSF), and $n \in \mathbb{R}^n$ is the noise vector. Here we simulate this process by filtering the true image x^{true} with a disk PSF of radius 15, causing the out-of-focus blurring, and then adding white Gaussian noise with standard deviation 0.01. We deal with 20 black/white images of a string of text, in the spirit of the Helsinki Deblur Challenge 2021[3].

Taking into account that the operator H is ill-conditioned and motivated by the features of the target image, one possible way of obtaining a valid estimate $\bar{x}$ is to minimize the following energy functional:

$$f(x) = \frac{1}{2}\|Hx - y\|^2 + \theta TV_\delta(x) + \eta B(x) + \iota_{x \in [0,1]^n}(x), \tag{67}$$

where:

- $TV_\delta(x) = \sum_{j=1}^n \sum_{s=1}^2 \sqrt{[\nabla_j x]_s^2 + \delta^2}$ is a smoothed version of the Total Variation functional for $\delta > 0$, being ∇_j the discrete gradient of x at pixel j;
- $B(x) = \sum_{j=1}^n x_j(1 - x_j)$ is a bimodal nonconvex function used to force the pixels to be either 1 or 0.

The term $TV_\delta(x)$ serves the purpose of inducing sparsity in the image gradient, in order to better preserve the edges of the letters, whereas the regularizer $B(x)$ induces sparsity in the image itself. The model parameters θ, η, δ were manually tuned for each of the algorithms tested, in order to provide a satisfactory quality of the restored image. The triplet (δ, θ, η) was set as $(0.38, 2.4 \times 10^{-4}, 2.9 \times 10^{-4})$ for IRVMILA, as $(0.61, 2.2 \times 10^{-4}, 9.6 \times 10^{-4})$ for iterative soft-thresholding algorithm (ISTA) and as $(0.35, 1.9 \times 10^{-4}, 4.7 \times 10^{-4})$ for VMILA. These values were chosen with a precision of two significant figures, through a two-step grid search: first we looked for the triplet $(\delta, \theta, \eta) \in \{0.1; 0.2; \ldots; 1\} \times \{10^{-4}; 2 \times 10^{-4}; \ldots; 10^{-3}\}^2$ that led to the best peak signal-to-noise ratio (PSNR), then we turned to a finer grid that had these as the middle points in each of the three dimensions.

---

[3] https://fips.fi/HDC2021.php.

The function in (67) can be cast in the form (21) by setting $f_0(x) = \frac{1}{2}\|Hx - y\|^2 + \theta TV_\delta(x)$, and $f_1(x) = \eta B(x) + \iota_{x\in[0,1]^n}(x) = \sum_{j=1}^{n+1}\phi_j(\psi_j(x)) + \chi(x)$, where:

$$\begin{cases} \phi_j(t) = \eta t(1-t), & j = 1, \ldots, n \\ \psi_j(x) = x_j, & j = 1, \ldots, n \\ \chi(x) = \iota_{x\in[0,1]^n}(x). \end{cases}$$

Furthermore, it is easy to see that f is coercive (see the appendix) and assumption 1 is satisfied. The weights of the convex majorant in (25) are defined by,

$$\omega_j^{(k)} = \eta(1 - 2x_j^{(k)}), \qquad j = 1, \ldots, n.$$

If we denote with $\omega^{(k)} = (\omega_1^{(k)}, \ldots, \omega_j^{(k)}, \ldots, \omega_n^{(k)})^T$, then the proximal operator required in algorithm 1 has the following explicit formulation:

$$\begin{aligned} \text{prox}_{\alpha_{k,i}\ell^{(k)}}^{D_{k,i}}(z) &= \underset{x\in\mathbb{R}^n}{\text{argmin}}\ \ell^{(k)}(x) + \frac{1}{2\alpha_{k,i}}\|x - z\|_{D_{k,i}}^2 \\ &= \underset{x\in[0,1]^n}{\text{argmin}}\ \sum_{j=1}^n \omega_j^{(k)}\psi_j(x) + \frac{1}{2\alpha_{k,i}}\|x - z\|_{D_{k,i}}^2 \\ &= \underset{x\in[0,1]^n}{\text{argmin}}\ (\omega^{(k)})^T x + \frac{1}{2\alpha_{k,i}}(x - z)^T D_{k,i}(x - z) \\ &= \min\left\{1, \max\left\{0, D_{k,i}^{-1}\left(D_{k,i}z - \alpha_{k,i}\omega^{(k)}\right)\right\}\right\}. \end{aligned} \qquad (68)$$

We tested the proposed IR-VMILA algorithm using two different strategies for the selection of the inner steplengths: (i) computing $\alpha_{k,i}$ by alternating the Barzilai–Borwein (BB) rules [29] according to the adaptive criterion employed in [37, section 3.2]; (ii) computing $\alpha_{k,i}$ as a pre-fixed, constant steplength. Moreover, we investigated the effect of the scaling matrix $D_{k,i}$ when chosen according to the split-gradient strategy, see e.g. [30, 37]. The variants of IR-VMILA equipped with the BB rules will be denoted as IR-VMILA BB (no scaling) and IR-VMILA BBS (with scaling). We compared IR-VMILA with the standard ISTA [6] and the VMILAn scheme presented in [4, 9]. For these algorithms, we chose a different splitting for the function f, gathering all the smooth term together in $f_0(x) = \frac{1}{2}\|Hx - y\|^2 + \theta TV_\delta(x) + \eta B(x)$, so that the resolvent operator needed in the backward step simply consists in the orthogonal projection onto the cartesian product $[0,1]^n$. For all algorithms, the outer iterations were stopped when either a maximum number of iterations $k_{\max} = 1000$ was achieved, or when the relative difference $e_r^{(k)} = |f(x^{(k)}) - f(x^{(k-1)})|/|f(x^{(k)})|$ was lower than $\epsilon = 10^{-7}$. Likewise, the inner iterations were halted when either a maximum number $I_{\max} = 25$ was reached, or when the relative difference of the convex majorant $f_0 + \ell^{(k)}$ value between two consecutive inner iterations was smaller than the tolerance ε. All of the algorithms were initialized at the blurred and noisy data y. The plots that are shown in the following are obtained by considering only the image in figure 1, whereas the results in table 1 are averaged across the 20 images considered.

In figure 2 we compare the performances of IR-VMILA BB and IR-VMILA BBS. For this problem, the use of the scaling matrix in combination with the BB rules does not seem to provide any considerable acceleration. On the other hand, in figure 3 we illustrate how the BB rules are crucial in speeding up the algorithm, as choosing constant steplengths leads to a run time more than 10 times greater, while still obtaining a similar quality in the reconstruction. The
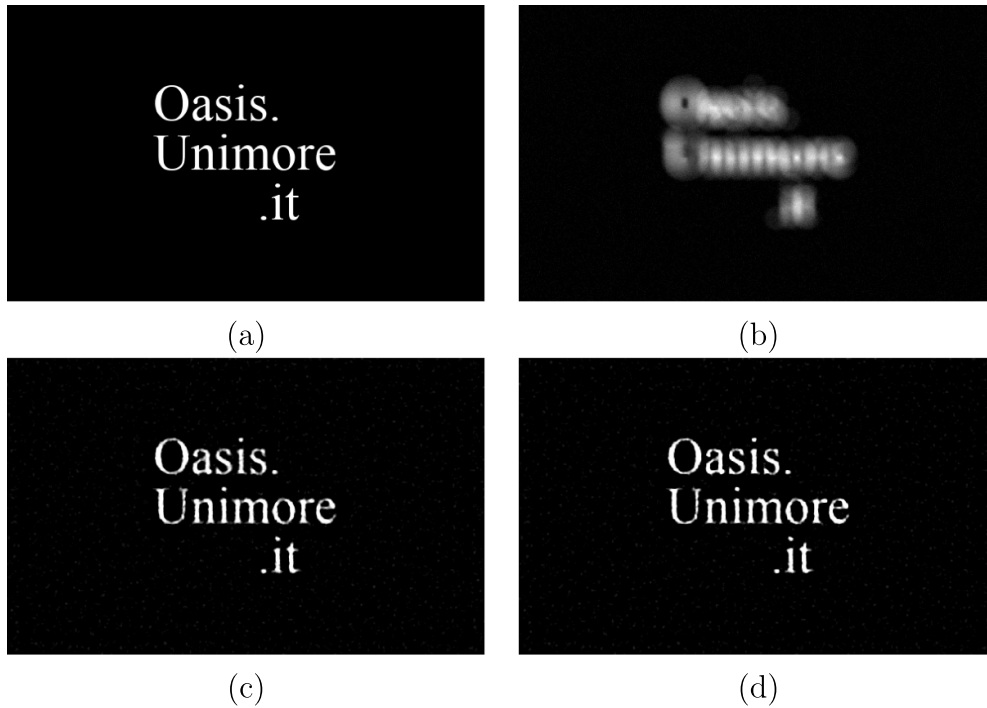
Figure 1. (a) Ground truth; (b) blurred sum of the observations; (c) and (d) reconstruction obtained by the proposed approach, using the BB rules for the steplength selection and the constant step $\alpha_k \equiv 1$ respectively.

Table 1. Performance evaluation for all the algorithms tested for the image deblurring problem. In the last column, an asterisk denotes that the maximum number of iterations was reached, while the number in parentheses indicates the average number of inner iterations (only for IR-VMILA).

	Avg. Time (s)	Avg. PSNR	Avg. Iters
IR-VMILA BB	**10.9564**	24.8211	**9.05 (22.3)**
IR-VMILA BBS	18.208	25.0229	13.25 (22.76)
IR-VMILA $\alpha_{k,i} \equiv 1$	281.333	**25.1481**	231.3 (24.42)
IR-VMILA $\alpha_{k,i} \equiv 1/L$	288.429	25.147	233.35 (24.43)
ISTA	63.0875	23.8714	1000*
ISTA w/ Backtracking	51.0974	20.8482	1000*
VMILAn	12.1967	21.4687	126.8

same, but with a different magnitude, can be said when IR-VMILA BB is compared against ISTA and VMILAn. In order to evaluate the quality of the local minimum found, we also report in table 1 the PSNR for each estimate $\bar{x}$. From this table, we observe that IR-VMILA yields a better image quality than the other competitors.
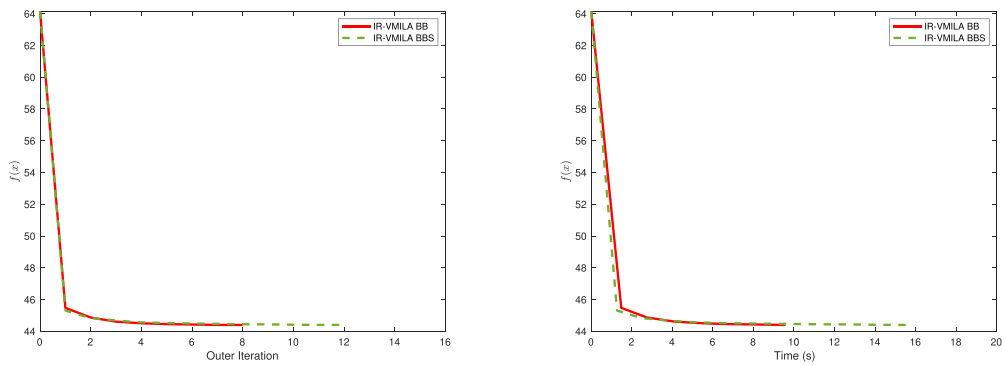
Figure 2. Decrease of the objective function with respect to the outer iterations (left) and computational time (right). Comparison between IR-VMILA equipped with the BB rules, with scaling (BBS) and without scaling (BB).
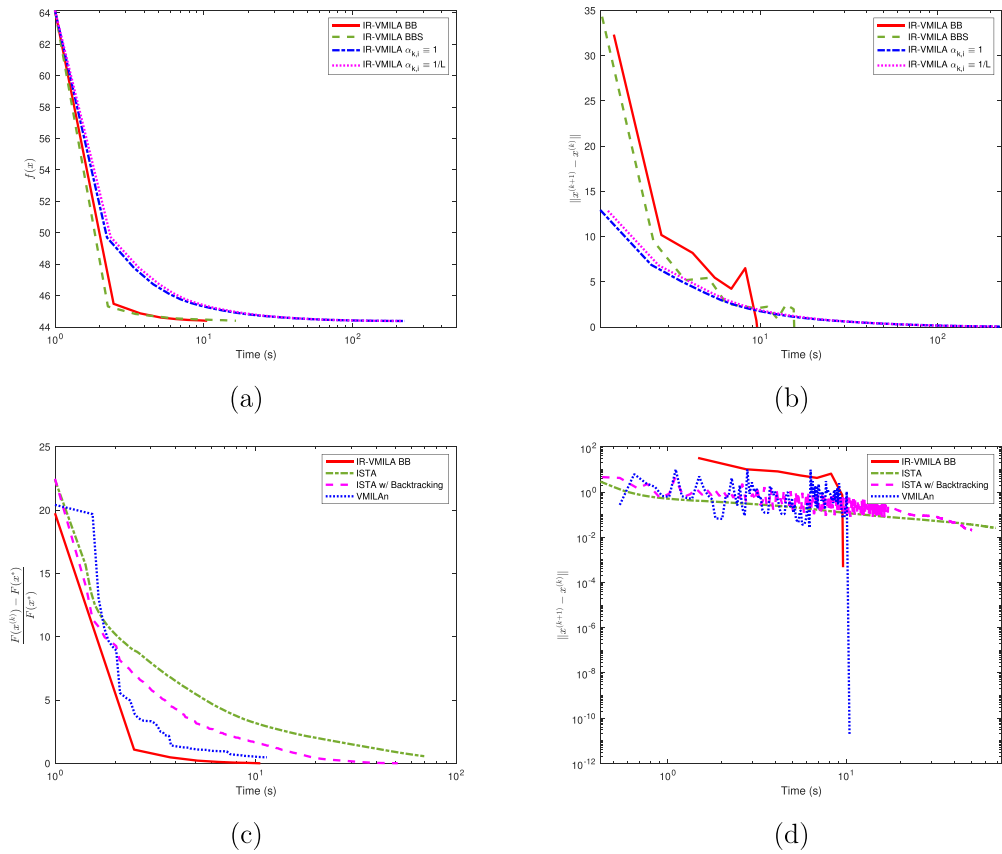
Figure 3. (a), (b) Comparison of several variants of our proposed algorithm: IR-VMILA BB (BB rules, no scaling), IR-VMILA (BB rules, scaling), IR-VMILA $\alpha_{k,i} = 1$ (constant steplength, no scaling), and IR-VMILA $\alpha_{k,i} = 1/L$ (constant steplength, no scaling); L denotes an estimate of the Lipschitz constant of ∇f_0; (c), (d) comparison between IR-VMILA BB, ISTA, ISTA with backtracking, and VMILAn ($F(x^*)$ is obtained by running each algorithm for 5000 iterations).

4.2. Image deblurring and denoising using nonconvex edge preserving regularization

In this section we illustrate the applicability of our approach to problems where the proximity operator of the function $\ell^{(k)}$ is not available in closed form. To this end, we consider a deblurring problem where the data $y \in \mathbb{R}^n$ is blurred and corrupted by Gaussian noise. Then, we seek for an approximation of the true image by minimizing the following functional [23]:

$$f(x) = \frac{1}{2}\|Hx - y\|^2 + \eta \sum_{j=1}^n \log\left(1 + \sqrt{[\nabla_j x]_1^2 + [\nabla_j x]_2^2}\right) + \iota_{x \geqslant 0}(x), \tag{69}$$

where $H \in \mathbb{R}^{n \times n}$ is the blurring operator and $\eta > 0$. As explained in [23], the second term in the above expression is motivated by the analysis of natural image statistics and, in general, it promotes solutions that are mainly smooth with some sharp discontinuities corresponding to edges. The objective function (69) is a special instance of (21), where $f_0(x) = \frac{1}{2}\|Hx - y\|^2$ and $f_1(x) = \eta \sum_{j=1}^n \log\left(1 + \sqrt{[\nabla_j x]_1^2 + [\nabla_j x]_2^2}\right) + \iota_{\geqslant 0}(x)$, with:

$$\begin{cases} \phi_j(t) = \eta \log(1 + t), & j = 1, \ldots, n \\ \psi_j(x) = \sqrt{[\nabla_j x]_1^2 + [\nabla_j x]_2^2}, & j = 1, \ldots, n \\ \chi(x) = \iota_{x \geqslant 0}(x). \end{cases}$$

It is easy to check that assumption 1 is satisfied. The weights of the convex majorant in (25) are defined by:

$$\omega_j^{(k)} = \frac{\eta}{1 + \psi_j(x^{(k)})}, \qquad j = 1, \ldots, n. \tag{70}$$

The proximal operator involved at the inner iterations of algorithm 1 is then defined as follows:

$$\text{prox}_{\alpha_{k,i}\ell^{(k)}}^{D_{k,i}}(z) = \underset{x \geqslant 0}{\text{argmin}} \sum_{j=1}^n \omega_j^{(k)} \sqrt{[\nabla_j x]_1^2 + [\nabla_j x]_2^2} + \frac{1}{2\alpha_{k,i}}(x - z)^T D_{k,i}(x - z). \tag{71}$$

The solution of the above minimization problem can not be expressed in explicit form. Therefore, within our convergence framework, the proximal gradient point $\hat{y}^{(k,i)} = \text{prox}_{\alpha_{k,i}\ell^{(k)}}^{D_{k,i}}(\tilde{x}^{(k,i)} - \alpha_{k,i}D_{k,i}^{-1}\nabla f_0(\tilde{x}^{(k,i)}))$ must be approximated by a point $\tilde{y}^{(k,i)}$ satisfying inequality (28). Such a point can be computed by applying an iterative optimization method to the dual of problem (71), until the primal-dual gap satisfies a suitable, easy to check condition [9, section 3.1]. Overall, when applied to the minimization of (69), algorithm 1 consists of three nested loops.

We define a test problem by selecting a natural image x^{true} whose pixels range is $[0, 1]$, convolving it with a Gaussian kernel (standard deviation 1.5), and then adding Gaussian noise of zero mean and standard deviation 0.01. The target image and its noisy version are depicted in figures 4(a) and (b). As for the model parameters, we empirically select $\eta = 5 \times 10^{-4}$, which produced a good visual quality for the reconstructions. The scaling matrices in algorithm 1 are set as $D_{k,i} = \mathcal{I}_n$, for all k and i, whereas the steplengths $\alpha_{k,i}$ are selected with the same alternation of the BB rules described in the previous section. The inner stopping condition is defined as described in the previous section with $\epsilon = 10^{-7}$ and $I_{\max} = 10$. The dual problem for approximating the proximal gradient point at each inner iteration is solved with FISTA, until condition (28) is satisfied with $\tau = 2 \times 10^{-6}$.

For comparison, we consider the method IRL1 [23, algorithm 3], implemented as described in the paper. More specifically, the inner convex problem is approximately solved by using
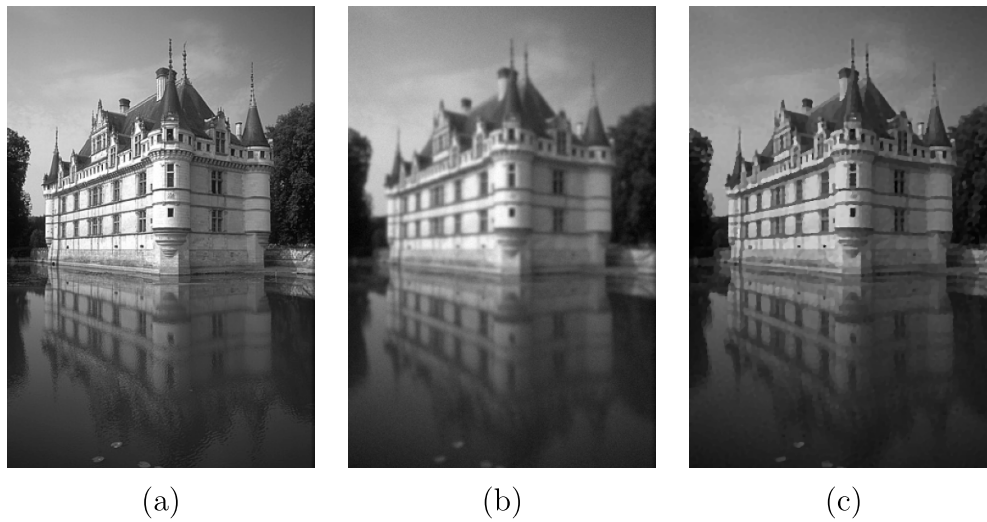
Figure 4. (a) ground truth; (b) noisy image (PSNR = 24.85); (c) restoration (PSNR = 26.73).
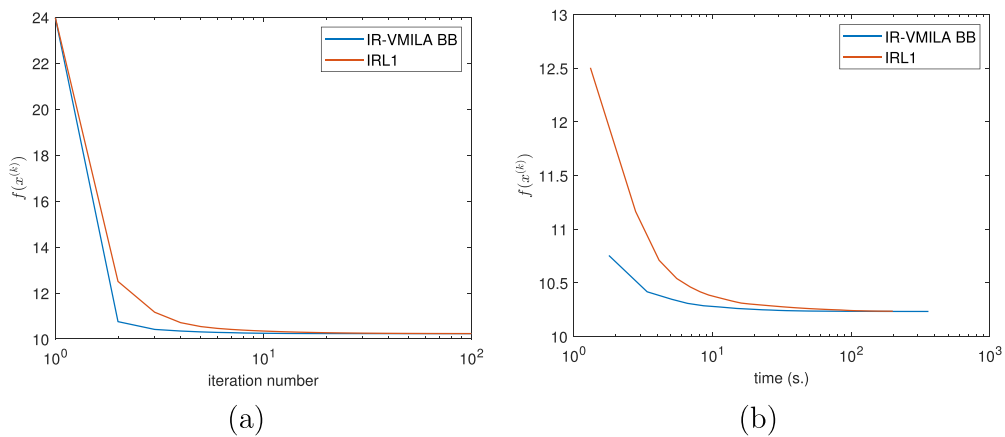
Figure 5. Numerical comparison: decrease of the objective function in logarithmic scale on the horizontal axis with respect to the iteration number (a) and the computational time (b).

the Chambolle-Pock method, which is stopped after 10 inner iterations. Both methods are initialized with the noisy blurred image *y* and run for 100 outer iterations.

The restored image obtained at the final iteration of IR-VMILA is reported in figure 4(c), while in figure 5 we report the objective function decrease for IR-VMILA and IRL1 with respect to the iteration number (a) and the computational time (b). From the plot we can conclude that IR-VMILA provides an accelerated convergence rate towards its limit point, which is mainly due to the variable steplength selection. The acceleration is more evident with respect to the iteration number, but it is still significant also in terms of the computational time. In this

regard, it should be noted that the implementation of IRL1 was tailored for this specific problem, whereas the IR-VMILA code was not optimized likewise.

4.3. Image super-resolution with CEL0

We now turn our attention to an interesting image super-resolution problem arising in the field of microscopy. The aim is to obtain an high resolution image $\bar{x} \in \mathbb{R}^{K^2 n}$ from a blurred, noisy and low resolution observation $y \in \mathbb{R}^n$, where $K > 0$ is the super-resolution factor. This is a common task in microscopy, where the physical limitations of the acquisition tools often lead to under-sampled images, causing close and small objects to remain undetected.

Recent super-resolution techniques in SMLM, like [38–40], propose to acquire an elevated number of under-sampled images, where each time only a limited random number of molecules is activated. After these acquisitions are cleaned of the noise and blur, they can be stacked to form the super-resolved image. The effectiveness of this process is influenced by the number of such acquisitions, which in turn is related to the density of the activated molecules. In order not to damage the object of interest, the volume of acquisitions should be restricted, leading to a high density of the molecules, risking again to have smaller objects to go unnoticed. With everything considered, using a sparsity inducing model when restoring the samples is a valid strategy to overcome this issue.

Here, specifically, we focus on the microscopy problem from the ISBI13 challenge[4]. The ground-truth image $x^{true} \in \mathbb{R}^{K^2 n}$, with $K = 4$ and $n = 64^2$, shows 8 slim tubular structures which were localized, using the scheme previously described, through 361 acquisitions $y_i \in \mathbb{R}^n$, $i = 1, \ldots, 361$, that are 4 times smaller than the true image.

If y represents one low resolution image corrupted by additive Gaussian noise and blur, then the corresponding super-resolution image $\bar{x}$ can be obtained as the minimum of the following energy functional:

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + R_{CEL0}(x) + \iota_{x \geqslant 0}(x),$$

where $A = SH$ is the composition of two operators: $H \in \mathbb{R}^{K^2 n \times K^2 n}$ is the blurring operator that represents the convolution with the PSF h, while $S \in \mathbb{R}^{n \times K^2 n}$ is the downsampling operator, which reduces each dimension of the image by a factor K. The CEL0 regularization term $R_{CEL0}(x)$ is meant to induce sparsity on the reconstruction by being an (exact) continuous approximation of the ℓ_0 pseudonorm [16]. Denoting with A_j the jth column of the operator A, the regularization term presents itself as:

$$R_{\text{CEL0}}(x) = \sum_{j=1}^{n} \varphi(|x_j|; \|A_j\|^2, \eta),$$

where

$$\varphi(t; A_j, \eta) = \eta - \frac{\|A_j\|^2}{2}\left(t - \frac{\sqrt{2\eta}}{\|A_j\|}\right)^2 \mathbb{1}_{\left\{t \leqslant \frac{\sqrt{2\eta}}{\|A_j\|}\right\}}.$$

In this case we set $f_0(x) = \frac{1}{2}\|Ax - y\|^2$ and $f_1(x) = R_{\text{CEL0}}(x) + \iota_{x \geqslant 0}(x)$. Taking into account the non-negativity constraint that is enforced in the problem through the indicator function,

---

[4] https://srm.epfl.ch/Datasets.

we can actually ignore the absolute values in the first argument of the function φ. Thus, the objective function can be framed in problem (21) by setting:

$$
\begin{cases}
\phi_j(t) = \varphi(t; A_j, \eta), & j = 1, \ldots, n \\
\psi_j(x) = x_j, & j = 1, \ldots, n \\
\chi(x) = \iota_{x \geqslant 0}(x).
\end{cases}
$$

Note that assumption 1 is satisfied; particularly, the coercivity of f holds, since the operator A has nonnegative entries and no identically null rows and columns (see the appendix). Consequently, the weights are given by:

$$
\omega_j^{(k)} = -\|A_j\|^2 \left(x_j^{(k)} - \frac{\sqrt{2\eta}}{\|A_j\|} \right) \mathbb{1}_{\left\{ x_j^{(k)} \leqslant \frac{\sqrt{2\eta}}{\|A_j\|} \right\}}, \qquad j = 1, \ldots, n,
$$

and the proximal operator of $\ell^{(k)}$, whose formulation only depends on the expression of the ψ_j, has the same form of equation (68).

We employed IR-VMILA to solve the resulting minimization problem and we compared the results with the ones obtained by running ISTA, the C2FB algorithm from [27] as well as the IRL1 algorithm [23] with FISTA as the inner solver, following the footsteps of the authors of [16]. As in the previous experiment, we equipped IR-VMILA with different choices for the steplength: the alternated Barzilai-Borwein rules employed in [37], the constant steplength $\alpha_{k,i} = 1$, or $\alpha_{k,i} = 1/L$, where L is an approximation of the Lipschitz constant of ∇f_0. Regarding the scaling matrix, we computed $D_{k,i}$ using again the split-gradient strategy [30]. On the other hand, the scaling matrix for C2FB has been obtained with the procedure described in [3]. All the algorithms were stopped when either the maximum number of $k_{\max} = 1000$ iterations was reached, or when the relative error of the objective function in two consecutive iterates was less than $\epsilon = 6 \times 10^{-6}$. However, we adopted different settings for the inner routines of the two iteratively reweighted schemes. In particular, the maximum number of inner iterations for IR-VMILA was set to $I_{\max} = 3$, whereas the corresponding parameter in C2FB was fixed as $I_{\max} = 50$ and as $I_{\max} = 25$ for IRL1-FISTA. All three algorithms stopped the inner iterations when the relative difference in the function $f_0 + l_k$ was smaller than 10^{-7}. The reason why we adopted such different parameters $I_{\max}$ for the algorithm IR-VMILA, with respect to IRL1-FISTA and C2FB, is that the quality of the solutions provided by these two for a smaller upper bound on the inner iterations was much worse than the ones provided by the other method. Regarding the regularization parameter η, we used the value $\eta = 0.1$ that is optimal for the performance of ISTA. For all schemes the initial point was chosen as $x^{(0)} = A^T y$. The plots presented in figure 6 are obtained by applying the algorithms to only one of the 361 frames.

In figure 7 we illustrate how, this time, the scaling matrix does indeed provide a boost for the proposed algorithm in terms of a faster decrease of the objective function. It reduces the total number of iterations by a big enough margin to justify the computational cost for constructing the matrix, resulting in an overall reduced computational time. Figure 8 shows the effects of the BB rules on the performance of IR-VMILA. As expected, an adaptive strategy for the steplength selection improves the computational time, especially when compared to the small steplength provided by the inverse of the Lipschitz constant of ∇f_0. In conclusion, a clever choice of the steplength parameter, possibly in the presence of a scaling matrix, seems to accelerate the proposed approach towards a good approximation of the optimal value.

When compared against the other algorithms tested for the ISBI13 challenge, IR-VMILA is the best performing of the three in terms of decrease of the objective function. Indeed, in figure 9 we see that IR-VMILA reaches the stopping criterion for the outer loop in fewer iterations compared to the other schemes. Furthermore, even though the iterations of IR-VMILA

(a)   (b)   (c)

(d)   (e)   (f)
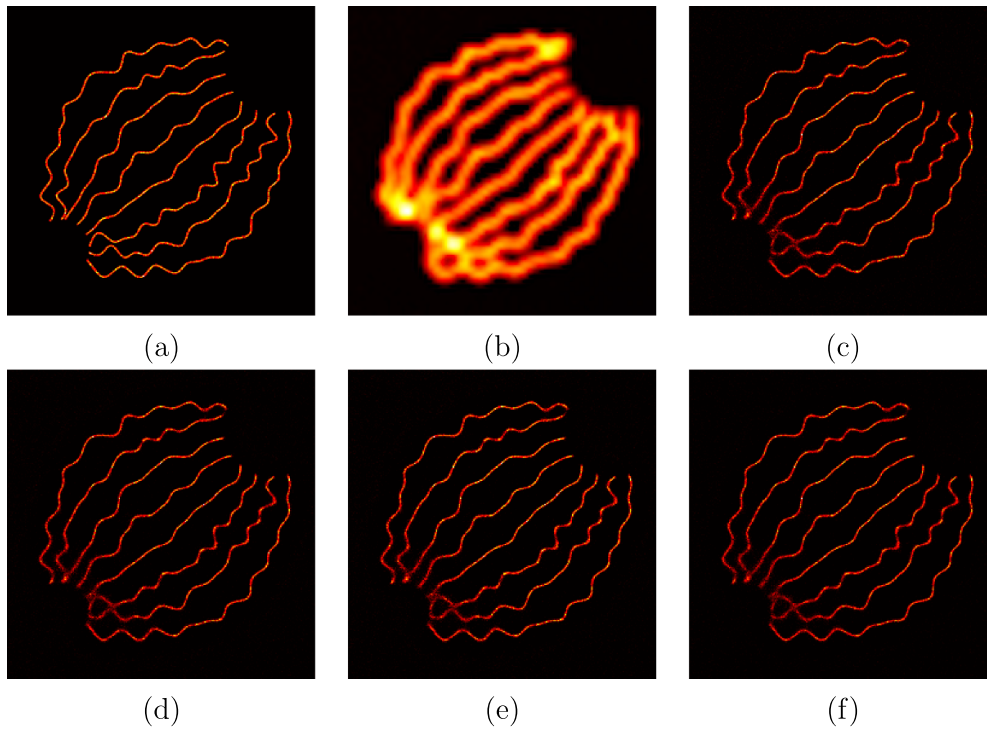
Figure 6. (a) Ground truth; (b) blurred sum of the observations; (c) reconstruction obtained by IR-VMILA equipped with the BB rules for the steplength selection; (d) reconstruction obtained with C2FB; (e) reconstruction obtained with ISTA; (f) reconstruction obtained with IRL1-FISTA.
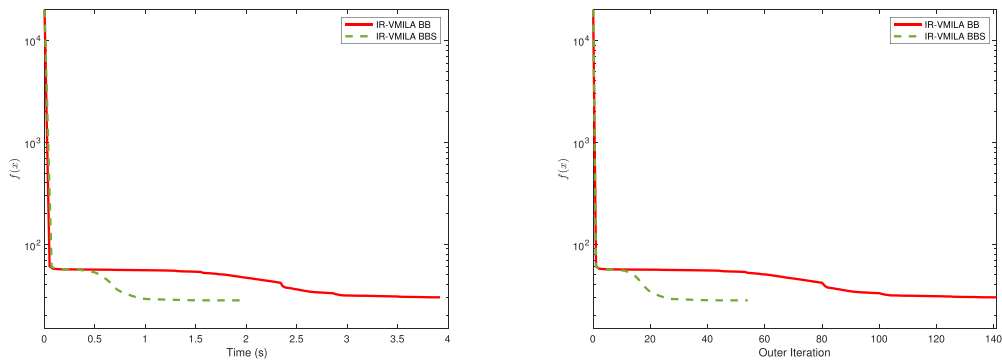
Figure 7. Decrease of the objective function with respect to computational time (left) and the outer iterations (right). Comparison between IR-VMILA equipped with the BB rules, with scaling (BBS) and without scaling (BB).
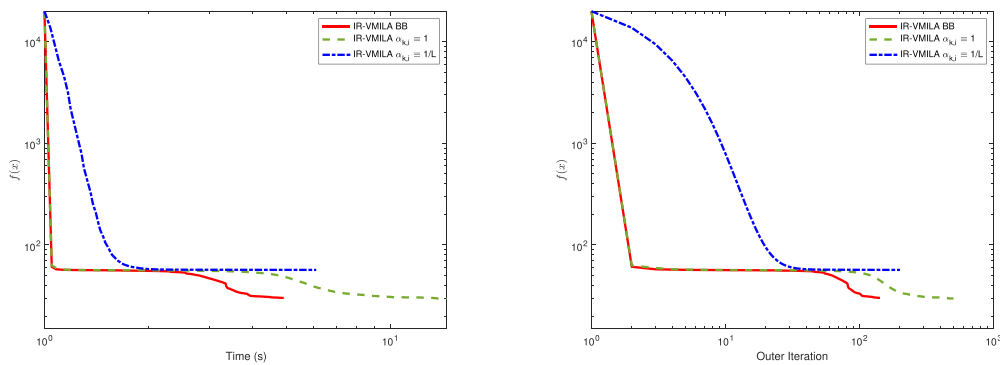
Figure 8. Decrease of the objective function with respect to computational time (left) and outer iterations (right). Comparison between IR-VMILA equipped with the BB rules (no scaling), IR-VMILA with $\alpha_{k,i} = 1$, and IR-VMILA with $\alpha_{k,i} = 1/L$, where L denotes an estimate of the Lipschitz constant of ∇f_0.

Table 2. Performance evaluation for the three tested schemes.

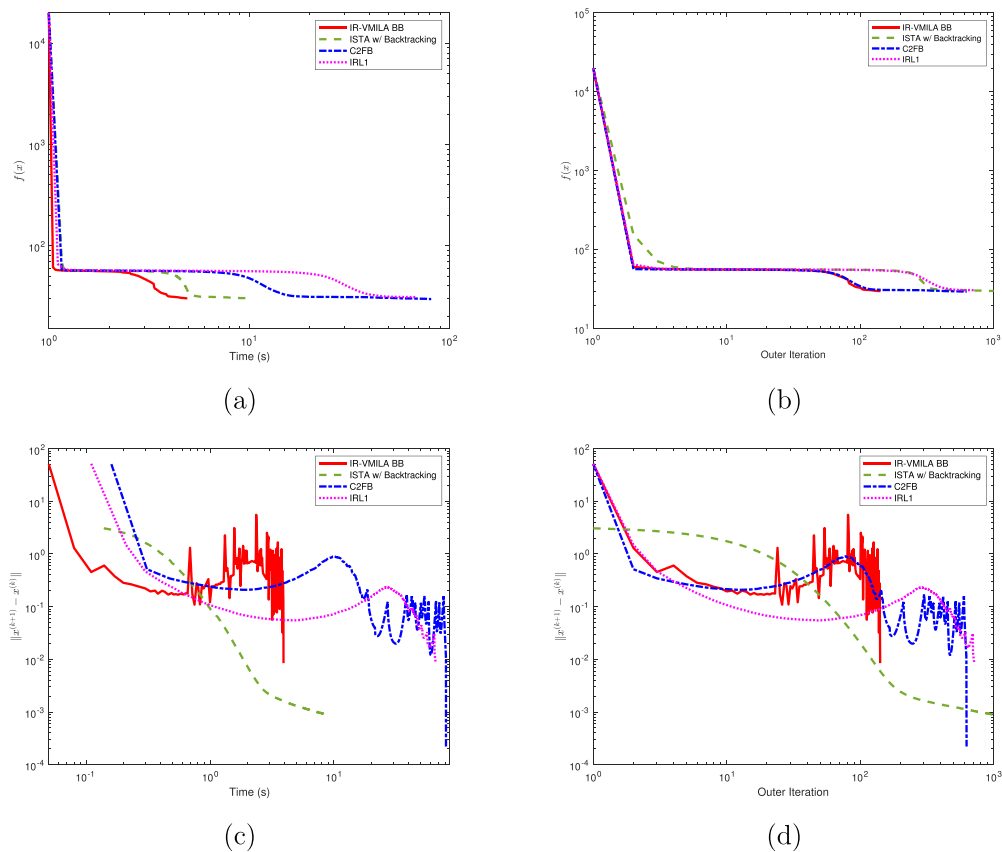|	IR-VMILA BB	ISTA w/ Backtracking	C2FB	IRL1-FISTA
J_0	0.0542	0.0545	0.0537	0.0545
J_2	0.4868	0.4891	0.4846	0.4918
J_4	0.5579	0.5633	0.5531	0.5861

are more costly than those of ISTA and IRL1-FISTA, the final computational time is still definitely in favor of the former.

In super-resolution microscopy, the goal is not to recover the correct intensity of the pixels, but rather the true position of the molecules. Thus, a sensible measure of the quality of the reconstruction is the Jaccard index, which is computed as

$$J_\delta = \frac{\#\text{of correct detections}}{(\#\text{of correct detections}) + (\#\text{of false negatives}) + (\#\text{of false positives})},$$

where $\delta \in \mathbb{N}$ is the number of pixels of tolerance. In table 2 we report the average Jaccard index measured for all the 361 reconstructions, for three different tolerances. Each column shows the Jaccard index relative to the reconstructions provided by one of the three algorithms IR-VMILA BB, ISTA with backtracking, C2FB, and IRL1-FISTA. We did not report the results for the IR-VMILA BBS variant, as we observed that the solution obtained with a nontrivial scaling matrix was worse than the nonscaled one, even if the objective function value was smaller. This could be due to the fact that the problem may have multiple local minima.

From table 2, we can observe that the three algorithms achieve comparable results in terms of accuracy, although IR-VMILA BB reconstructs the image with a significantly reduced amount of time, as figure 9 shows.

Figure 9. (a), (b) Decrease of the objective function with respect to computational time (left) and outer iterations (right). Comparison between IR-VMILA BB, ISTA with back-tracking, C2FB and IRL1-FISTA. (c), (d) Plot of the quantity $\|x^{(k+1)} - x^{(k)}\|$ over time (left) and over the iterations (right).

5. Conclusions

In this paper we have proposed a novel iteratively reweighted algorithm, where the solution of the inner subproblem is obtained by a finite number of linesearch based forward–backward iterations. In our approach, acceleration strategies related to the selection of the forward–backward parameters can be included. We provided the convergence analysis of the proposed algorithm and the results of a numerical experience on a couple of relevant problems. In particular, the novel algorithm seems to be promising as a tool for solving image super-resolution problems in microscopy. Future work will address the design of effective rules for the selection of the algorithmic parameters and for the specific microscopy application, as well as the definition of a suitable strategy to achieve maximum accuracy in the molecule localization while preserving the acceleration effect of the algorithm.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix

In this appendix, we show that the function:

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + R(x) + \iota_\Omega(x)$$

is coercive, where $\Omega = \{x \in \mathbb{R}^n : x \geqslant 0\}$ is the nonnegative orthant, $R : \mathbb{R}^n \to \mathbb{R}$ is a nonnegative function, and A is assumed to have nonnegative entries and no identically null rows and columns. Indeed, for all $x \in \Omega$, it is easy to write down the following inequalities:

$$
\begin{aligned}
f(x) &= \frac{1}{2}\|Ax - y\|^2 + R_{CEL0}(x) + \iota_{\geqslant 0}(x) \geqslant \frac{1}{2}\|Ax - y\|^2 \\
&= \frac{1}{2}\|Ax\|^2 + \frac{1}{2}\|y\|^2 - \langle y, Ax \rangle \\
&\geqslant \frac{1}{2}\|Ax\|^2 + \frac{1}{2}\|y\|^2 - \|y\|\|Ax\| \\
&= \|Ax\|\left(\frac{1}{2}\|Ax\| - \|b\|\right) + \frac{1}{2}\|y\|^2,
\end{aligned}
$$

where the first inequality follows from the definition of $\iota_{\geqslant 0}(x)$ and the nonnegativity of R_{CEL0}, whereas the third one from the Cauchy-Schwarz inequality. Let $\{x^{(\ell)}\}_{\ell \in \mathbb{N}} \subseteq \Omega$ be any sequence in the nonnegative orthant such that $\|x^{(\ell)}\| \xrightarrow[\ell \to +\infty]{} +\infty$. Define $j(\ell) = \underset{j=1,\dots,n}{\operatorname{argmax}}\, x_j^{(\ell)}$ for all $\ell \in \mathbb{N}$, then it follows that $|x_{j(\ell)}^{(\ell)}| \xrightarrow[\ell \to +\infty]{} +\infty$. Finally, let $i(\ell)$ be the first row such that the element $A_{i(\ell)j(\ell)}$ of the $j(\ell)$th column is positive. Then, focusing on $\|Ax^{(\ell)}\|$, we note that:

$$
\begin{aligned}
\|Ax^{(\ell)}\| &= \sqrt{\sum_{i=1}^n \left(\sum_{j=1}^n A_{ij}x_j^{(\ell)}\right)^2} \\
&\geqslant \sqrt{\sum_{i=1}^n (A_{ij(\ell)}x_{j(\ell)}^{(\ell)})^2} \geqslant A_{i(\ell)j(\ell)}|x_{j(\ell)}^{(\ell)}| \\
&\geqslant \left(\min_{m \in \mathbb{N}} A_{i(m)j(m)}\right) \cdot |x_{j(\ell)}^{(\ell)}| \xrightarrow[\ell \to +\infty]{} +\infty,
\end{aligned}
$$

where we used also the fact that the iterates $x^{(\ell)}$ and the matrix A have all nonnegative entries. Then $f(x^{(\ell)}) \xrightarrow[\ell \to +\infty]{} +\infty$ and hence f is coercive.

ORCID iDs

S Bonettini ⓘ https://orcid.org/0000-0003-2936-365X
D Pezzi ⓘ https://orcid.org/0000-0001-7396-9806
M Prato ⓘ https://orcid.org/0000-0002-7327-3347
S Rebegoldi ⓘ https://orcid.org/0000-0002-1232-5481

References

[1] Combettes P L and Wajs V R 2005 Signal recovery by proximal forward–backward splitting *Multiscale Model. Simul.* **4** 1168–200

[2] Combettes P L and Pesquet J-C 2011 Proximal splitting methods in signal processing *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (*Springer Optimization and Its Applications*) ed H H Bauschke, R S Burachik, P L Combettes, V Elser, D R Luke and H Wolkowicz (New York: Springer) pp 185–212

[3] Chouzenoux E, Pesquet J-C and Repetti A 2014 Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function *J. Optim. Theory Appl.* **162** 107–32

[4] Bonettini S, Loris I, Porta F, Prato M and Rebegoldi S 2017 On the convergence of a linesearch based proximal-gradient method for nonconvex optimization *Inverse Problems* **33** 055005

[5] Lee C-P and Wright S J 2019 Inexact successive quadratic approximation for regularized optimization *Comput. Optim. Appl.* **72** 641–74

[6] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202

[7] Ochs P, Chen Y, Brox T and Pock T 2014 iPiano: inertial proximal algorithm for non-convex optimization *SIAM J. Imaging Sci.* **7** 1388–419

[8] Attouch H, Bolte J and Svaiter B F 2013 Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting and regularized Gauss–Seidel methods *Math. Program.* **137** 91–129

[9] Bonettini S, Prato M and Rebegoldi S 2021 New convergence results for the inexact variable metric forward–backward method *Appl. Math. Comput.* **392** 125719

[10] Frankel P, Garrigos G and Peypouquet J 2015 Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates *J. Optim. Theory Appl.* **165** 874–900

[11] Ochs P 2019 Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano *SIAM J. Optim.* **29** 541–70

[12] Figueiredo M A T, Bioucas-Dias J M and Nowak R D 2007 Majorization–minimization algorithms for wavelet–based image restoration *IEEE Trans. Image Process.* **16** 2980–91

[13] Chen Y, Pock T, Ranftl R and Bischof H 2013 Revisiting loss-specific training of filter-based MRFs for image restoration *Pattern Recognition* ed J Weickert, M Hein and B Schiele (Berlin: Springer) pp 271–81

[14] Candés E J, Wakin M B and Boyd S P 2008 Enhancing sparsity by reweighted ℓ_1 minimization *J. Fourier Anal. Appl.* **14** 877–905

[15] Fan J and Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties *J. Am. Stat. Assoc.* **96** 1348–60

[16] Soubies E, Blanc-Féraud L and Aubert G 2015 A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem *SIAM J. Imaging Sci.* **8** 1607–39

[17] Gazagnes S, Soubies E and Blanc-Féraud L 2017 High density molecule localization for super-resolution microscopy using CEL0 based sparse approximation *2017 IEEE 14th Int. Symp. on Biomedical Imaging (ISBI 2017)* pp 28–31

[18] Chan R H, Lazzaro D, Morigi S and Sgallari F 2019 A non-convex nonseparable approach to single-molecule localization microscopy *Scale Space and Variational Methods in Computer Vision* ed J Lellmann, M Burger and J Modersitzki (Cham: Springer) pp 498–509

[19] Lazzaretti M, Calatroni L and Estatico C 2021 A continuous, non-convex & sparse super-resolution approach for fluorescence microscopy data with Poisson noise *2021 21st Int. Conf. on Computational Science and Its Applications (ICCSA)* pp 80–86

[20] Bolte J, Sabach S and Teboulle M 2014 Proximal alternating linearized minimization for nonconvex and nonsmooth problems *Math. Program.* **146** 459–94

[21] Chouzenoux E, Pesquet J-C and Repetti A 2016 A block coordinate variable metric forward-backward algorithm *J. Glob. Optim.* **66** 457–85

[22] Weiszfeld E 1937 Sur le point pour lequel la somme des distances de *n* points donnés est minimum *Tohoku Math. J.* **43** 355–86

[23] Ochs P, Dosovitskiy A, Brox T and Pock T 2015 On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision *SIAM J. Imaging Sci.* **8** 331–72

[24] Gorodnitsky I F and Rao B D 1997 Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm *IEEE Trans. Signal Process.* **45** 600–16

[25] Chen X and Zhou W 2014 Convergence of the reweighted ℓ_1 minimization algorithm for $\ell_2 - \ell_p$ minimization *Comput. Optim. Appl.* **59** 47–61

[26] Ochs P, Dosovitskiy A, Pock T and Brox T 2013 An iterated ℓ_1 algorithm for non-smooth non-convex optimization in computer vision *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR) (Portland, OR)* pp 1759–66

[27] Repetti A and Wiaux Y 2020 Variable metric forward–backward algorithm for composite minimization problems *SIAM J. Optim.* **31** 1215–41

[28] Lazzaretti M, Rebegoldi S, Calatroni L and Estatico C 2021 A scaled and adaptive FISTA algorithm for signal-dependent sparse image super-resolution problems *Scale Space and Variational Methods in Computer Vision* ed A Elmoataz, J Fadili, Y Quéau, J Rabin and L Simon (Cham: Springer) pp 242–53

[29] Barzilai J and Borwein J M 1988 Two-point step size gradient methods *IMA J. Numer. Anal.* **8** 141–8

[30] Bonettini S, Porta F, Ruggiero V and Zanni L 2021 Variable metric techniques for forward–backward methods in imaging *J. Comput. Appl. Math.* **385** 113192

[31] Rockafellar R T, Wets R J-B and Wets M 1998 *Variational Analysis* (*Grundlehren der Mathematischen Wissenschaften* vol 317) (Berlin: Springer)

[32] Rockafellar R T 1970 *Convex Analysis* (Princeton, NJ: Princeton University Press)

[33] Bolte J, Daniilidis A and Lewis A 2006 The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems *SIAM J. Optim.* **17** 1205–23

[34] Birgin E G, Martinez J M and Raydan M 2003 Inexact spectral projected gradient methods on convex sets *IMA J. Numer. Anal.* **23** 539–59

[35] Bonettini S, Loris I, Porta F and Prato M 2016 Variable metric inexact line–search based methods for nonsmooth optimization *SIAM J. Optim.* **26** 891–921

[36] Bonettini S, Prato M and Rebegoldi S 2020 Convergence of inexact forward-backward algorithms using the forward-backward envelope *SIAM J. Optim.* **30** 3069–97

[37] Zanella R, Boccacci P, Zanni L and Bertero M 2009 Efficient gradient projection methods for edge-preserving removal of Poisson noise *Inverse Problems* **25** 045010

[38] Betzig E, Patterson G H, Sougrat R, Lindwasser O W, Olenych S, Bonifacino J S, Davidson M W, Lippincott-Schwartz J and Hess H F 2006 Imaging intracellular fluorescent proteins at nanometer resolution *Science* **313** 1642–5

[39] Sage D, Kirshner H, Pengo T, Stuurman N, Min J, Manley S and Unser M 2015 Quantitative evaluation of software packages for single-molecule localization microscopy *Nat. Methods* **12** 717–24

[40] Rust M J, Bates M and Zhuang X 2006 Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM) *Nat. Methods* **3** 793–6