Nesterov's accelerated gradient method for nonlinear ill-posed problems with a locally convex residual functional

To cite this article: Simon Hubmer and Ronny Ramlau 2018 *Inverse Problems* **34** 095003

Nesterov's accelerated gradient method for nonlinear ill-posed problems with a locally convex residual functional

Simon Hubmer[1] and Ronny Ramlau[1,2]

[1] Johann Radon Institute Linz, Altenbergerstraße 69, A-4040 Linz, Austria
[2] Johannes Kepler University Linz, Institute of Industrial Mathematics, Altenbergerstraße 69, A-4040 Linz, Austria

E-mail: simon.hubmer@ricam.oeaw.ac.at, ronny.ramlau@jku.at and ronny.ramlau@ricam.oeaw.ac.at

Abstract

In this paper, we consider Nesterov's accelerated gradient method for solving nonlinear inverse and ill-posed problems. Known to be a fast gradient-based iterative method for solving well-posed convex optimization problems, this method also leads to promising results for ill-posed problems. Here, we provide convergence analysis of this method for ill-posed problems based on the assumption of a locally convex residual functional. Furthermore, we demonstrate the usefulness of the method on a number of numerical examples based on a nonlinear diagonal operator and on an inverse problem in auto-convolution.

Keywords: Nesterov's accelerated gradient method, Landweber iteration, two-point gradient method, regularization method, inverse and ill-posed problems, auto-convolution

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

1. Introduction

In this paper, consider nonlinear inverse problems of the form

$$F(x) = y, \tag{1.1}$$

where $F : \mathcal{D}(F) \subset \mathcal{X} \to \mathcal{Y}$ is a continuously Fréchet-differentiable, nonlinear operator between real Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Throughout this paper we assume that (1.1) has a solution x_*, which need not be unique. Furthermore, we assume that instead of y, we are only given noisy data y^δ satisfying

$$\left\| y - y^\delta \right\| \leqslant \delta. \tag{1.2}$$

Since we are interested in ill-posed problems, we need to use regularization methods in order to obtain stable approximations of solutions of (1.1). The two most prominent examples of such methods are *Tikhonov regularization* and *Landweber iteration*.

In *Tikhonov regularization*, one attempts to approximate an x_0-minimum-norm solution $x^\dagger$ of (1.1), i.e. a solution of $F(x) = y$ with minimal distance to a given initial guess x_0, by minimizing the functional

$$\mathcal{T}_\alpha^\delta(x) := \left\| F(x) - y^\delta \right\|^2 + \alpha \left\| x - x_0 \right\|^2, \tag{1.3}$$

where α is a suitably chosen regularization parameter. Under very mild assumptions on F, it can be shown that the minimizers of $\mathcal{T}_\alpha^\delta$, usually denoted by x_α^δ, converge subsequentially to a minimum norm solution $x^\dagger$ as $\delta \to 0$, given that α and the noise level δ are coupled in an appropriate way [9]. While for linear operators F the minimization of $\mathcal{T}_\alpha^\delta$ is straightforward, in the case of nonlinear operators F the computation of x_α^δ requires the global minimization of the then also nonlinear functional $\mathcal{T}_\alpha^\delta$, which is rather difficult and usually done using various iterative optimization algorithms.

This motivates the direct application of iterative algorithms for solving (1.1), the most popular of which being *Landweber iteration*, given by

$$\begin{aligned} x_{k+1}^\delta &= x_k^\delta + \omega F'(x_k^\delta)^*(y^\delta - F(x_k^\delta)), \\ x_0^\delta &= x_0, \end{aligned} \tag{1.4}$$

where ω is a scaling parameter and x_0 is again a given initial guess. Seen in the context of classical optimization algorithms, Landweber iteration is nothing else than the gradient descent method applied to the functional

$$\Phi^\delta(x) := \frac{1}{2} \left\| F(x) - y^\delta \right\|^2, \tag{1.5}$$

and therefore, in order to arrive at a convergent regularization method, one has to use a suitable stopping rule. In [9] it was shown that if one uses the discrepancy principle, i.e. stops the iteration after k_* steps, where k_* is the smallest integer such that

$$\left\| y^\delta - F(x_{k_*}^\delta) \right\| \leqslant \tau \delta < \left\| y^\delta - F(x_k^\delta) \right\|, \qquad 0 \leqslant k < k_*, \tag{1.6}$$

with a suitable constant $\tau > 1$, then Landweber iteration gives rise to a convergent regularization method, as long as some additional assumptions, most notably the (strong) tangential cone condition

$$\begin{aligned} \| F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x}) \| &\leqslant \eta \| F(x) - F(\tilde{x}) \|, \qquad \eta < \frac{1}{2} \\ x, \tilde{x} &\in \mathcal{B}_{2\rho}(x_0), \end{aligned} \tag{1.7}$$

where $\mathcal{B}_{2\rho}(x_0)$ denotes the closed ball of radius 2ρ around x_0, is satisfied. Since condition (1.7) poses strong restrictions on the nonlinearity of F which are not always satisfied, attempts have been made to use weaker conditions instead [32]. For example, assuming only the weak tangential cone condition

$$\langle F(x) - F(x_*) - F'(x)(x - x_*), F(x) - F(x_*) \rangle \leqslant \eta \, \|F(x) - F(x_*)\|^2$$
$$\forall x \in \mathcal{B}_\rho(x_0), \qquad 0 < \eta < 1, \tag{1.8}$$

with x_* denoting a solution of (1.1) to hold, one can show weak convergence of Landweber iteration [32]. Similarly, if the residual functional $\Phi^0(x)$ defined by (1.5) is (locally) convex, weak subsequential convergence of the iterates of Landweber iteration to a stationary point of $\Phi^0(x)$ can be proven. Even though they both lead to convergence in the weak topology, besides some results presented in [32], the connections between the local convexity of the residual functional and the (weak) tangential cone condition remain largely unexplored. In his recent paper [24], Kindermann showed that both the local convexity of the residual functional and the weak tangential cone condition imply another condition, which he termed $NC(0, \beta > 0)$, and which is sufficient to guarantee weak subsequential convergence of the iterates.

As is well known, Landweber iteration is quite slow [23]. Hence, acceleration strategies have to be used in order to speed it up and make it applicable in practise. Acceleration methods and their analysis for linear problems can be found for example in [9] and [13]. Unfortunately, since their convergence proofs are mainly based on spectral theory, their analysis cannot be generalized to nonlinear problems immediately. However, there are some acceleration strategies for Landweber iteration for nonlinear ill-posed problems, for example [26, 30].

As an alternative to (accelerated) Landweber-type methods, one could think of using second order iterative methods for solving (1.1), such as the Levenberg–Marquardt method [14, 20]

$$x_{k+1}^\delta = x_k^\delta + (F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I)^{-1} F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \tag{1.9}$$

or the iteratively regularized Gauss–Newton method [6, 22]

$$x_{k+1}^\delta = x_k^\delta + (F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I)^{-1} (F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)) + \alpha_k(x_0 - x_k^\delta)). \tag{1.10}$$

The advantage of those methods [23] is that they require much less iterations to meet their respective stopping criteria compared to Landweber iteration or the steepest descent method. However, each update step of those iterations might take considerably longer than one step of Landweber iteration, due to the fact that in both cases a linear system involving the operator

$$F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I$$

has to be solved. In practical applications, this usually means that a huge linear system of equations has to be solved, which often proves to be costly, if not infeasible. Hence, accelerated Landweber type methods avoiding this drawback are desirable in practise.

In case that the residual functional $\Phi^\delta(x)$ is locally convex, one could think of using methods from convex optimization to minimize $\Phi^\delta(x)$, instead of using the gradient method like in Landweber iteration. One of those methods, which works remarkably well for nonlinear, convex and well-posed optimization problems of the form

$$\min\{\Phi(x) \,|\, x \in \mathcal{X}\} \tag{1.11}$$

was first introduced by Nesterov in [25] and is given by

$$z_k = x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}),$$
$$x_{k+1} = z_k - \omega(\nabla\Phi(z_k)), \tag{1.12}$$

where again ω is a given scaling parameter and $\alpha \geqslant 3$ (with $\alpha = 3$ being common practise). This so-called *Nesterov acceleration scheme* is of particular interest, since not only is it extremely easy to implement, but Nesterov himself was also able to prove that it generates a sequence of iterates x_k for which there holds

$$\|\Phi(x_k) - \Phi(x_*)\| = \mathcal{O}(k^{-2}), \tag{1.13}$$

where x_* is any solution of (1.11). This is a big improvement over the classical rate $\mathcal{O}(k^{-1})$. The even further improved rate $o(k^{-2})$ for $\alpha > 3$ was recently proven in [2].

Furthermore, Nesterov's acceleration scheme can also be used to solve compound optimization problems of the form

$$\min\{\Phi(x) + \Psi(x) \,|\, x \in \mathcal{X}\}, \tag{1.14}$$

where both $\Phi(x)$ and $\Psi(x)$ are convex functionals, and is in this case given by

$$z_k = x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\omega\Psi}\left(z_k - \omega(\nabla\Phi(z_k))\right), \tag{1.15}$$

where the proximal operator $\mathrm{prox}_{\omega\Psi}(.)$ is defined by

$$\mathrm{prox}_{\omega\Psi}(x) := \arg\min_u\left\{\omega\Psi(u) + \frac{1}{2}\|x - u\|^2\right\}. \tag{1.16}$$

If in addition to being convex, Ψ is proper and lower-semicontinous and Φ is continuously Fréchet differentiable with a Lipschitz continuous gradient, then it was again shown in [2] that the sequence defined by (1.15) satisfies

$$\|(\Phi + \Psi)(x_k) - (\Phi + \Psi)(x_*)\| = \mathcal{O}(k^{-2}), \tag{1.17}$$

or even $o(k^{-2})$ if $\alpha > 3$, which is again much faster than ordinary first order methods for minimizing (1.14). This accelerating property was exploited in the highly successful FISTA algorithm [4], designed for the fast solution of linear ill-posed problems with sparsity constraints. Since for linear operators the residual functional Φ^δ is globally convex, minimizing the resulting Tikhonov functional (1.3) exactly fits into the category of minimization problems considered in (1.15).

Motivated by the above considerations, one could think of applying Nesterov's acceleration scheme (1.12) to the residual functional Φ^δ, which leads to the algorithm

$$z_k^\delta = x_k^\delta + \frac{k-1}{k+\alpha-1}(x_k^\delta - x_{k-1}^\delta),$$
$$x_{k+1}^\delta = z_k^\delta + \omega F'(z_k^\delta)^*(y^\delta - F(z_k^\delta)),$$
$$x_0^\delta = x_{-1}^\delta = x_0. \tag{1.18}$$

In case that the operator F is linear, Neubauer showed in [28] that, combined with a suitable stopping rule and under a source condition, (1.18) gives rise to a convergent regularization method and that convergence rates can be obtained. Furthermore, the authors of [17] showed that certain generalizations of Nesterov's acceleration scheme, termed *two-point gradient (TPG) methods* and given by

$$
\begin{aligned}
z_k^\delta &= x_k^\delta + \lambda_k^\delta(x_k^\delta - x_{k-1}^\delta), \\
x_{k+1}^\delta &= z_k^\delta + \alpha_k^\delta s_k^\delta, \qquad s_k^\delta := F'(z_k^\delta)^*(y^\delta - F(z_k^\delta)), \\
x_0^\delta &= x_{-1}^\delta = x_0,
\end{aligned}
\tag{1.19}
$$

give rise to convergent regularization methods, as long as the tangential cone condition (1.7) is satisfied and the stepsizes α_k^δ and the combination parameters λ_k^δ are coupled in a suitable way. However, the convergence analysis of the methods (1.19) does not cover the choice

$$
\lambda_k^\delta = \frac{k-1}{k+\alpha-1},
\tag{1.20}
$$

i.e. the choice originally proposed by Nesterov and the one which shows by far the best results numerically [17, 18, 21]. The main reason for this is that the techniques employed there works with the monotonicity of the iteration, i.e. the iterate x_{k+1}^δ always has to be a better approximation of the solution x_* than x_k^δ, which is not necessarily satisfied for the choice (1.20).

The key ingredient for proving the fast rates (1.13) and (1.17) is the convexity of the residual functional $\Phi(x)$. Since, except for linear operators, we cannot hope that this holds globally, we assume that $\Phi^0(x)$, i.e. the functional $\Phi^\delta(x)$ defined by (1.5) with exact data $y = y^\delta$, corresponding to $\delta = 0$, is convex in a neighbourhood of the initial guess. This neighbourhood has to be sufficiently large encompassing the sought solution x_*, or equivalently, the initial guess x_0 has to be sufficiently close to the solution x_*. Assuming that $F(x) = y$ has a solution x_* in $\mathcal{B}_\rho(x_0)$, where now and in the following, $\mathcal{B}_\rho(x_0)$ denotes the closed ball with radius ρ around x_0, the key assumption is that Φ^0 is convex in $\mathcal{B}_{6\rho}(x_0)$. As mentioned before, Nesterov's acceleration scheme yields a non-monotonous sequence of iterates, which might possible leave the ball $\mathcal{B}_{6\rho}(x_0)$. However, by assumption the sought for solution x_* lies in the ball $\mathcal{B}_\rho(x_0)$. Hence, defining the functional

$$
\Psi(x) := \begin{cases} 0, & x \in \mathcal{B}_{2\rho}(x_0), \\ \infty, & x \notin \mathcal{B}_{2\rho}(x_0), \end{cases}
\tag{1.21}
$$

we can, instead of using (1.12), which would lead to algorithm (1.18), use (1.15), noting that still the fast rate (1.17) can be expected for $\delta = 0$. This leads to the algorithm

$$
\begin{aligned}
z_k^\delta &= x_k^\delta + \frac{k-1}{k+\alpha-1}(x_k^\delta - x_{k-1}^\delta), \\
x_{k+1}^\delta &= \mathrm{prox}_{\omega\Psi}\left(z_k^\delta + \omega F'(z_k^\delta)^*(y^\delta - F(z_k^\delta))\right), \\
x_0^\delta &= x_{-1}^\delta = x_0,
\end{aligned}
\tag{1.22}
$$

which we consider throughout this paper.

2. Convergence analysis I

In this section we provide a convergence analysis of Nesterov's accelerated gradient method (1.22). Concerning notation, whenever we consider the noise-free case $y = y^\delta$ corresponding to $\delta = 0$, we replace δ by 0 in all variables depending on δ, e.g. we write Φ^0 instead of Φ^δ. For carrying out the analysis, we have to make a set of assumptions, already indicated in the introduction.

Assumption 2.1. *Let ρ be a positive number such that $\mathcal{B}_{6\rho}(x_0) \subset \mathcal{D}(F)$.*

1. *The operator $F : \mathcal{D}(F) \subset \mathcal{X} \to \mathcal{Y}$ is continuously Fréchet differentiable between the real Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with inner products $\langle .,. \rangle$ and norms $\|.\|$. Furthermore, let F be weakly sequentially closed on $\mathcal{B}_{2\rho}(x_0)$.*
2. *The equation $F(x) = y$ has a solution $x_* \in \mathcal{B}_\rho(x_0)$.*
3. *The data y^δ satisfies $\left\| y - y^\delta \right\| \leqslant \delta$.*
4. *The functional Φ^0 defined by (1.5) with $\delta = 0$ is convex and has a Lipschitz continuous gradient $\nabla \Phi^0$ with Lipschitz constant L on $\mathcal{B}_{6\rho}(x_0)$, i.e.*

$$\Phi^0(\lambda x_1 + (1-\lambda)x_2) \leqslant \lambda \Phi^0(x_1) + (1-\lambda)\Phi^0(x_2), \qquad \forall x_1, x_2 \in \mathcal{B}_{6\rho}(x_0), \tag{2.1}$$

$$\left\| \nabla \Phi^0(x_1) - \nabla \Phi^0(x_2) \right\| \leqslant L \left\| x_1 - x_2 \right\|, \qquad \forall x_1, x_2 \in \mathcal{B}_{6\rho}(x_0). \tag{2.2}$$

5. *For α in (1.22) there holds $\alpha > 3$ and the scaling parameter ω satisfies $0 < \omega < \frac{1}{L}$.*

Note that since $\mathcal{B}_{2\rho}(x_0)$ is weakly closed and given the continuity of F, a sufficient condition for the weak sequential closedness assumption to hold is that F is compact.

We now turn to the convergence analysis of Nesterov's accelerated gradient method (1.22). Throughout this analysis, if not explicitly stated otherwise, assumption 2.1 is in force. Note first that from F being continuously Fréchet differentiable, we can derive that there exists an $\bar{\omega}$ such that

$$\|F'(x)\| \leqslant \bar{\omega}, \qquad \forall x \in \mathcal{B}_{6\rho}(x_0). \tag{2.3}$$

Next, note that since $\mathcal{B}_{2\rho}(x)$ denotes a closed ball around x, the functional Ψ, in addition to being proper and convex, is also lower-semicontinuous, an assumption required in the proofs in [2], which we need in various places of this paper. Furthermore, it immediately follows from the definition (1.16) of the proximal operator $\mathrm{prox}_{\omega \Psi}(.)$ that

$$\mathrm{prox}_{\omega \Psi}(x) = \arg\min_{u \in \mathcal{X}} \left\{ \omega \Psi(u) + \frac{1}{2} \|x - u\|^2 \right\} = \operatorname*{argmin}_{u \in \mathcal{B}_{2\rho}(x_0)} \left\{ \frac{1}{2} \|x - u\|^2 \right\}, \tag{2.4}$$

since Ψ defined by (1.21) is equal to ∞ outside $\mathcal{B}_{2\rho}(x_0)$. Hence, since obviously $\mathcal{B}_{2\rho}(x_0)$ is a convex set, $\mathrm{prox}_{\omega \Psi}(.)$ is nothing else than the metric projection onto $\mathcal{B}_{2\rho}(x_0)$, and is therefore Lipschitz continuous with Lipschitz constant smaller or equal to 1. Consequently, given an estimate of ρ, the implementation of $\mathrm{prox}_{\omega \Psi}(.)$ is exceedingly simple in this setting, and therefore, one iteration step of (1.22) and (1.4) require roughly the same amount of computational effort.

Finally, note that due to the convexity of Φ^0, the set S defined by

$$\mathcal{S} := \left\{ x \in \mathcal{B}_{2\rho}(x_0) \,|\, F(x) = y \right\}, \tag{2.5}$$

is a convex subset of $\mathcal{B}_{2\rho}(x_0)$ and hence, there exists a unique x_0-minimum-norm solution $x^\dagger$, which is defined by

$$x^\dagger := \arg\min_{x \in \mathcal{S}} \|x - x_0\|, \tag{2.6}$$

which is nothing else than the orthogonal projection of x_0 onto the set $\mathcal{S}$.

The following convergence analysis is largely based on the ideas of the paper [2] of Attouch and Peypouquet, which we reference from frequently throughout this analysis. Following their arguments, we start by making the following

Definition 2.1. For Φ^δ and Ψ defined by (1.5) and (1.21), we define

$$\Theta^\delta(x) := \Phi^\delta(x) + \Psi(x) = \frac{1}{2}\left\|F(x) - y^\delta\right\|^2 + \begin{cases} 0, & x \in \mathcal{B}_{2\rho}(x_0), \\ \infty, & x \notin \mathcal{B}_{2\rho}(x_0). \end{cases} \quad (2.7)$$

The energy functional $\mathcal{E}^\delta$ is defined by

$$\mathcal{E}^\delta(k) := \frac{2\omega}{\alpha - 1}(k + \alpha - 2)^2(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*)) + (\alpha - 1)\left\|w_k^\delta - x_*\right\|^2, \quad (2.8)$$

where the sequence w_k^δ is defined by

$$w_k^\delta := \frac{k + \alpha - 1}{\alpha - 1}z_k^\delta - \frac{k}{\alpha - 1}x_k^\delta = x_k^\delta + \frac{k - 1}{\alpha - 1}\left(x_k^\delta - x_{k-1}^\delta\right). \quad (2.9)$$

Furthermore, we introduce the operator $G_\omega^\delta : \mathcal{D}(F) \subset \mathcal{X} \to \mathcal{Y}$, given by

$$G_\omega^\delta(x) := \frac{1}{\omega}\left(x - \text{prox}_{\omega\Psi}\left(x - \omega\nabla\Phi^\delta(x)\right)\right). \quad (2.10)$$

Using definition 2.1, we can now write the update step for x_{k+1}^δ in the form

$$x_{k+1}^\delta = z_k^\delta - \omega\, G_\omega^\delta(z_k^\delta),$$

from which, together with (2.9) and the definition of z_k^δ, it follows that

$$\begin{aligned} w_{k+1}^\delta &= \frac{k + \alpha}{\alpha - 1}\left(x_{k+1}^\delta + \frac{k}{k + \alpha}\left(x_{k+1}^\delta - x_k^\delta\right)\right) - \frac{k + 1}{\alpha - 1}x_{k+1}^\delta = \frac{k + \alpha - 1}{\alpha - 1}x_{k+1}^\delta - \frac{k}{\alpha - 1}x_k^\delta \\ &= \frac{k + \alpha - 1}{\alpha - 1}\left(z_k^\delta - \omega\, G_\omega^\delta(z_k^\delta)\right) - \frac{k}{\alpha - 1}x_k^\delta = w_k^\delta - \frac{\omega}{\alpha - 1}(k + \alpha - 1)G_\omega^\delta(z_k^\delta). \end{aligned} \quad (2.11)$$

As a first result, we show that both z_k^δ and x_k^δ stay within $\mathcal{B}_{6\rho}(x_0)$ during the iteration.

Lemma 2.1. *Under the assumption 2.1, the sequence of iterates x_k^δ and z_k^δ defined by (1.22) is well-defined. Furthermore, $x_k^\delta \in \mathcal{B}_{2\rho}(x_0)$ and $z_k^\delta \in \mathcal{B}_{6\rho}(x_0)$ for all $k \in \mathbb{N}$.*

Proof. This follows by induction from $x_0^\delta = x_{-1}^\delta = x_0 \in \mathcal{B}_\rho(x_0)$, the observation

$$\begin{aligned} \left\|z_k^\delta - x_0\right\| &\leqslant (1 + \frac{k - 1}{k + \alpha - 1})\left\|x_k^\delta - x_0\right\| + \frac{k - 1}{k + \alpha - 1}\left\|x_{k-1}^\delta - x_0\right\| \\ &\leqslant 2\left\|x_k^\delta - x_0\right\| + \left\|x_{k-1}^\delta - x_0\right\|, \end{aligned}$$

and the fact that by the definition of $\text{prox}_{\omega\Psi}(x)$, x_k^δ is always an element of $\mathcal{B}_{2\rho}(x_0)$. $\qquad\square$

Since the functional Θ^0 is assumed to be convex in $\mathcal{B}_{6\rho}(x_0)$, we can deduce:

Lemma 2.2. *Under assumption 2.1, for all $x, z \in \mathcal{B}_{6\rho}(x_0)$ there holds*

$$\Theta^0(z - \omega G_\omega^0(z)) \leqslant \Theta^0(x) + \left\langle G_\omega^0(z), z - x\right\rangle - \frac{\omega}{2}\left\|G_\omega^0(z)\right\|^2.$$

Proof. This lemma is also used in [2]. However, the sources for it cited there do not exactly cover our setting with Φ^δ being defined on $\mathcal{D}(F) \subset \mathcal{X}$ only. Hence, we here give an elementary proof of the assertion. Note first that due to the Lipschitz continuity of Φ^0 in $\mathcal{B}_{6\rho}(x_0)$ and

the fact that $\omega < 1/L$ we have

$$\Phi^0(u) \leqslant \Phi^0(v) + \left\langle \nabla\Phi^0(v), u - v \right\rangle + \frac{1}{2\omega} \|u - v\|^2, \qquad \forall u, v \in \mathcal{B}_{6\rho}(x_0).$$

Now since Φ^0 is convex on $\mathcal{B}_{6\rho}(x_0)$, also have [3]

$$\Phi^0(v) + \left\langle \nabla\Phi^0(v), w - v \right\rangle \leqslant \Phi^0(w), \qquad \forall v, w \in \mathcal{B}_{6\rho}(x_0),$$

and therefore, combining the above two inequalities, we get

$$\Phi^0(u) \leqslant \Phi^0(w) + \left\langle \nabla\Phi^0(v), u - w \right\rangle + \frac{1}{2\omega} \|u - v\|^2, \qquad \forall u, v, w \in \mathcal{B}_{6\rho}(x_0).$$

Using this result for $u = z - \omega G_\omega^0(z)$, $v = z$, $w = x$, noting that for $x, z \in \mathcal{B}_{6\rho}(x_0)$ there holds $u, v, w \in \mathcal{B}_{6\rho}(x_0)$, we get

$$\Phi^0(z - \omega G_\omega^0(z)) \leqslant \Phi^0(x) + \left\langle \nabla\Phi^0(z), z - \omega G_\omega^0(z) - x \right\rangle + \frac{\omega}{2} \left\| G_\omega^0(z) \right\|^2. \tag{2.12}$$

Next, note that since $z - \omega G_\omega^0(z) = \mathrm{prox}_{\omega\Psi}\left(z - \omega\nabla\Phi^0(z)\right)$, a standard result from proximal operator theory [3, proposition 12.26] implies that there holds

$$\begin{aligned}
\Psi(z - \omega G_\omega^0(z)) &\leqslant \Psi(x) + \frac{1}{\omega}\left\langle (z - \omega G_\omega^0(z)) - x, (z - \omega\nabla\Phi^0(z)) - (z - \omega G_\omega^0(z)) \right\rangle \\
&= \Psi(x) + \left\langle z - \omega G_\omega^0(z) - x, -\nabla\Phi^0(z) + G_\omega^0(z) \right\rangle \\
&= \Psi(x) - \left\langle z - \omega G_\omega^0(z) - x, \nabla\Phi^0(z) \right\rangle + \left\langle z - x, G_\omega^0(z) \right\rangle - \omega\left\| G_\omega^0(z) \right\|^2.
\end{aligned}$$

Adding this inequality to (2.12) and using the fact that by definition $\Theta^0 = \Phi^0 + \Psi$ immediately yields the assertion. $\qquad\square$

    We want to derive a similar inequality also for the functionals Θ^δ. The following lemma is of vital importance for doing that:

Lemma 2.3. *Let assumption 2.1 hold, let $x, z \in \mathcal{B}_{6\rho}(x_0)$ and define*

$$\begin{aligned}
R_1 &:= \Theta^0(z - \omega G_\omega^\delta(z)) - \Theta^0(z - \omega G_\omega^0(z)), \\
R_2 &:= \Theta^\delta(z - \omega G_\omega^\delta(z)) - \Theta^0(z - \omega G_\omega^\delta(z)), \\
R_3 &:= \Theta^0(x) - \Theta^\delta(x), \quad R_4 := \left\langle G_\omega^0(z) - G_\omega^\delta(z), z - x \right\rangle, \\
R_5 &:= \frac{\omega}{2}\left(\left\| G_\omega^\delta(z) \right\|^2 - \left\| G_\omega^0(z) \right\|^2 \right),
\end{aligned} \tag{2.13}$$

as well as

$$R := R_1 + R_2 + R_3 + R_4 + R_5. \tag{2.14}$$

Then there holds

$$\Theta^\delta(z - \omega G_\omega^\delta(z)) \leqslant \Theta^\delta(x) + \left\langle G_\omega^\delta(z), z - x \right\rangle - \frac{\omega}{2}\left\| G_\omega^\delta(z) \right\|^2 + R.$$

Proof.   Using lemma 2.2 we get

$$\Theta^\delta(z - \omega G_\omega^\delta(z)) = \Theta^0(z - \omega G_\omega^0(z)) + R_1 + R_2$$

$$\leqslant \Theta^0(x) + \left\langle G_\omega^0(z), z - x \right\rangle - \frac{\omega}{2} \left\| G_\omega^0(z) \right\|^2 + R_1 + R_2$$

$$= \Theta^\delta(x) + \left\langle G_\omega^\delta(z), z - x \right\rangle - \frac{\omega}{2} \left\| G_\omega^\delta(z) \right\|^2 + R_1 + R_2 + R_3 + R_4 + R_5,$$

from which the statement of the theorem immediately follows. $\square$

Next, we show that the R_i and hence, also R, can be bounded in terms of $\delta + \delta^2$.

Proposition 2.4.   *Let assumption 2.1 hold, let $x \in \mathcal{B}_{2\rho}(x_0)$ and $z \in \mathcal{B}_{6\rho}(x_0)$ and let the $R_1, \ldots, R_5$ be defined by (2.14). Then, with $\bar\omega$ as in (2.3), there holds*

$$R_1 \leqslant \frac{1}{2} \bar\omega^4 \omega^2 \delta^2 + 2 \bar\omega^3 \omega \rho \delta,$$

$$R_2 \leqslant \frac{3}{2} \delta^2 + 2\rho \, \bar\omega \, \delta,$$

$$R_3 \leqslant \frac{3}{2} \delta^2 + 2\rho \, \bar\omega \, \delta,$$

$$R_4 \leqslant 8\rho \bar\omega \delta,$$

$$R_5 \leqslant \frac{\omega}{2} \bar\omega^2 \, \delta^2 + 8\rho \, \bar\omega \, \delta.$$

Proof.   The following somewhat long but elementary proof uses mainly the boundedness and Lipschitz continuity assumptions made above. For the following, let $x \in \mathcal{B}_{2\rho}(x_0)$ and $z \in \mathcal{B}_{6\rho}(x_0)$. We treat each of the R_i terms separately, starting with

$$R_1 = \Theta^0(z - \omega G_\omega^\delta(z)) - \Theta^0(z - \omega G_\omega^0(z))$$

$$= \frac{1}{2} \left\| F(z - \omega G_\omega^\delta(z)) - y \right\|^2 - \frac{1}{2} \left\| F(z - \omega G_\omega^0(z)) - y \right\|^2$$

$$= \frac{1}{2} \left\| F(z - \omega G_\omega^\delta(z)) - F(z - \omega G_\omega^0(z)) \right\|^2$$

$$\quad - \left\langle F(z - \omega G_\omega^\delta(z)) - F(z - \omega G_\omega^0(z)), F(z - \omega G_\omega^0(z)) - y \right\rangle$$

$$\leqslant \frac{1}{2} \left\| F(z - \omega G_\omega^\delta(z)) - F(z - \omega G_\omega^0(z)) \right\|^2$$

$$\quad + \left\| F(z - \omega G_\omega^\delta(z)) - F(z - \omega G_\omega^0(z)) \right\| \left\| F(z - \omega G_\omega^0(z)) - y \right\|.$$

Since we have

$$\left\| F(z - \omega G_\omega^\delta(z)) - F(z - \omega G_\omega^0(z)) \right\| \leqslant \bar\omega \left\| \omega G_\omega^\delta(z) - \omega G_\omega^0(z) \right\|$$

$$\leqslant \bar\omega \left\| \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^\delta(z) \right) - \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi(z) \right) \right\|$$

$$\leqslant \bar\omega \, \omega \left\| \nabla \Phi^\delta(z) - \nabla \Phi(z) \right\|$$

$$= \bar\omega \, \omega \left\| F'(z)^*(y - y^\delta) \right\| \leqslant \bar\omega^2 \, \omega \left\| y - y^\delta \right\| \leqslant \bar\omega^2 \, \omega \, \delta,$$

and

$$\left\| F(z - \omega G_\omega^0(z)) - y \right\| \leqslant \bar{\omega} \left\| \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^\delta(z) \right) - x_* \right\| \leqslant 2\rho\,\bar{\omega},$$

there holds

$$R_1 \leqslant \frac{1}{2} (\bar{\omega}^2\,\omega\,\delta)^2 + (\bar{\omega}^2\,\omega\,\delta)2\rho\,\bar{\omega} = \left(\frac{1}{2}\bar{\omega}^4\omega^2 \right) \delta^2 + \left(2\bar{\omega}^3\omega\rho \right) \delta.$$

Next, we look at

$$
\begin{aligned}
R_2 &= \Theta^\delta(z - \omega G_\omega^\delta(z)) - \Theta^0(z - \omega G_\omega^\delta(z)) \\
&= \frac{1}{2} \left\| y - y^\delta \right\|^2 + \left\langle F(z - \omega G_\omega^\delta(z)) - y^\delta, y - y^\delta \right\rangle \\
&= \frac{3}{2} \left\| y - y^\delta \right\|^2 + \left\langle F(z - \omega G_\omega^\delta(z)) - y, y - y^\delta \right\rangle \\
&\leqslant \frac{3}{2}\delta^2 + \left\| F(z - \omega G_\omega^\delta(z)) - F(x_*) \right\| \delta \\
&\leqslant \frac{3}{2}\delta^2 + 2\rho\,\bar{\omega}\,\delta.
\end{aligned}
$$

Similarly to above, for the next term we get

$$
\begin{aligned}
R_3 &= \Theta^0(x) - \Theta^\delta(x) = \frac{1}{2} \left\| F(x) - y \right\|^2 - \frac{1}{2} \left\| F(x) - y^\delta \right\|^2 \\
&= \frac{1}{2} \left\| y - y^\delta \right\|^2 + \left\langle F(x) - y^\delta, y - y^\delta \right\rangle \\
&= \frac{3}{2} \left\| y - y^\delta \right\|^2 + \left\langle F(x) - y, y - y^\delta \right\rangle \\
&\leqslant \frac{3}{2}\delta^2 + \left\| F(x) - F(x_*) \right\| \delta \\
&\leqslant \frac{3}{2}\delta^2 + 2\rho\,\bar{\omega}\,\delta.
\end{aligned}
$$

Furthermore, together with the Lipschitz continuity of $\mathrm{prox}_{\omega\Psi}(.)$, we get

$$
\begin{aligned}
R_4 &= \left\langle G_\omega^0(z) - G_\omega^\delta(z), z - x \right\rangle \\
&= \frac{1}{\omega} \left\langle \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^\delta(z) \right) - \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^0(z) \right), z - x \right\rangle \\
&\leqslant \frac{1}{\omega} \left\| \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^\delta(z) \right) - \mathrm{prox}_{\omega\Psi} \left(z - \omega \nabla \Phi^0(z) \right) \right\| \left\| z - x \right\| \\
&\leqslant \left\| \nabla \Phi^\delta(z) - \nabla \Phi^0(z) \right\| \left\| z - x \right\| \leqslant 8\rho \left\| F'(z)(y - y^\delta) \right\| \leqslant 8\rho\,\bar{\omega}\delta.
\end{aligned}
$$

Finally, for the last term, we get

$$
\begin{aligned}
R_5 &= \frac{\omega}{2} \left(\left\| G_\omega^\delta(z) \right\|^2 - \left\| G_\omega^0(z) \right\|^2 \right) \\
&= \frac{\omega}{2} \left\| G_\omega^\delta(z) - G_\omega^0(z) \right\|^2 + \omega \left\langle G_\omega^\delta(z) - G_\omega^0(z), G_\omega^0(z) \right\rangle \\
&\leqslant \frac{\omega}{2} \left\| G_\omega^\delta(z) - G_\omega^0(z) \right\|^2 + \omega \left\| G_\omega^\delta(z) - G_\omega^0(z) \right\| \left\| G_\omega^0(z) \right\| \\
&\leqslant \frac{\omega}{2}\bar{\omega}^2\,\delta^2 + \omega\,\bar{\omega}\,\delta \left\| G_\omega^0(z) \right\| \leqslant \frac{\omega}{2}\bar{\omega}^2\,\delta^2 + 8\rho\,\bar{\omega}\,\delta,
\end{aligned}
$$

which concludes the proof.   □

As an immediate consequence, we get the following:

Corollary 2.5.   *Let assumption 2.1 hold and let $x, z \in \mathcal{B}_{6\rho}(x_0)$. If we define*

$$
\begin{aligned}
c_1 &= 2\bar{\omega}^3\,\omega\,\rho + 20\rho\,\bar{\omega}, \\
c_2 &= 3 + \frac{1}{2}\bar{\omega}^4\omega^2 + \frac{1}{2}\omega\bar{\omega}^2,
\end{aligned}
\tag{2.15}
$$

then there holds

$$
\Theta^\delta(z - \omega G_\omega^\delta(z)) \leqslant \Theta^\delta(x) + \left\langle G_\omega^\delta(z), z - x \right\rangle - \frac{\omega}{2}\left\| G_\omega^\delta(z) \right\|^2 + c_1\delta + c_2\delta^2.
$$

Proof.   This immediately follows from lemma 2.2 and proposition 2.4.   □

Combining the above, we are now able to arrive at the following important result:

Proposition 2.6.   *Let assumption 2.1 hold, let the sequence of iterates x_k^δ and z_k^δ be given by (1.22) and let c_1 and c_2 be defined by (2.15). If we define*

$$
\Delta(\delta) := c_1\delta + c_2\delta^2,
\tag{2.16}
$$

then there holds

$$
\Theta^\delta(z_k^\delta - \omega G_\omega^\delta(z_k^\delta)) \leqslant \Theta^\delta(x_k^\delta) + \left\langle G_\omega^\delta(z_k^\delta), z_k^\delta - x_k^\delta \right\rangle - \frac{\omega}{2}\left\| G_\omega^\delta(z_k^\delta) \right\|^2 + \Delta(\delta),
\tag{2.17}
$$

$$
\Theta^\delta(z_k^\delta - \omega G_\omega^\delta(z_k^\delta)) \leqslant \Theta^\delta(x_*) + \left\langle G_\omega^\delta(z_k^\delta), z_k^\delta - x_* \right\rangle - \frac{\omega}{2}\left\| G_\omega^\delta(z_k^\delta) \right\|^2 + \Delta(\delta).
\tag{2.18}
$$

Proof.   This immediately follows from lemma 2.1 and corollary 2.5.   □

Using the above proposition, we are now able to derive the important

Theorem 2.7.   *Let assumption 2.1 hold and let the sequence of iterates x_k^δ and z_k^δ be given by (1.22) and let $\Delta(\delta)$ be defined by (2.16). Then there holds*

$$
\mathcal{E}^\delta(k+1) + \frac{2\omega}{\alpha - 1}\left(k(\alpha - 3)\left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) - (k + \alpha - 1)^2\Delta(\delta) \right) \leqslant \mathcal{E}^\delta(k).
\tag{2.19}
$$

Proof.   This proof is adapted from the corresponding result in [2], the difference being the term $\Delta(\delta)$. We start by multiplying inequality (2.17) by $\frac{k}{k+\alpha-1}$ and inequality (2.18) by $\frac{\alpha-1}{k+\alpha-1}$. Adding the results and using the fact that $x_{k+1}^\delta = z_k^\delta - \omega G_\omega^\delta(z_k^\delta)$, we get

$$
\begin{aligned}
\Theta^\delta(x_{k+1}^\delta) \leqslant {} & \frac{k}{k+\alpha-1}\Theta^\delta(x_k^\delta) + \frac{\alpha-1}{k+\alpha-1}\Theta^\delta(x_*) - \frac{\omega}{2}\left\| G_\omega^\delta(z_k^\delta) \right\|^2 + \Delta(\delta) \\
& + \left\langle G_\omega^\delta(z_k^\delta), \frac{k}{k+\alpha-1}(z_k^\delta - x_k^\delta) + \frac{\alpha-1}{k+\alpha-1}(z_k^\delta - x_*) \right\rangle.
\end{aligned}
$$

Since

$$\frac{k}{k+\alpha-1}(z_k^\delta - x_k^\delta) + \frac{\alpha-1}{k+\alpha-1}(z_k^\delta - x_*) = \frac{\alpha-1}{k+\alpha-1}(w_k^\delta - x_*),$$

we obtain

$$\Theta^\delta(x_{k+1}^\delta) \leqslant \frac{k}{k+\alpha-1}\Theta^\delta(x_k^\delta) + \frac{\alpha-1}{k+\alpha-1}\Theta^\delta(x_*) - \frac{\omega}{2}\left\|G_\omega^\delta(z_k^\delta)\right\|^2 + \Delta(\delta)$$
$$\frac{\alpha-1}{k+\alpha-1}\left\langle G_\omega^\delta(z_k^\delta), w_k^\delta - x_* \right\rangle. \tag{2.20}$$

Next, observe that it follows from (2.11) that

$$w_{k+1}^\delta - x_* = w_k^\delta - x_* - \frac{\omega}{\alpha-1}(k+\alpha-1)G_\omega^\delta(z_k^\delta).$$

After developing

$$\left\|w_{k+1}^\delta - x_*\right\|^2 = \left\|w_k^\delta - x_*\right\|^2 - 2\frac{\omega}{\alpha-1}(k+\alpha-1)\left\langle w_k^\delta - x_*, G_\omega^\delta(z_k^\delta) \right\rangle$$
$$+ \frac{\omega^2}{(\alpha-1)^2}(k+\alpha-1)^2 \left\|G_\omega^\delta(z_k^\delta)\right\|^2,$$

and multiplying the above expression by $\frac{(\alpha-1)^2}{2\omega(k+\alpha-1)^2}$, we get

$$\frac{(\alpha-1)^2}{2\omega(k+\alpha-1)^2}\left(\left\|w_k^\delta - x_*\right\|^2 - \left\|w_{k+1}^\delta - x_*\right\|^2\right)$$
$$= \frac{\alpha-1}{k+\alpha-1}\left\langle G_\omega^\delta(z_k^\delta), w_k^\delta - x_* \right\rangle - \frac{\omega}{2}\left\|G_\omega^\delta(z_k^\delta)\right\|^2.$$

Replacing this in inequality (2.20) above, we get

$$\Theta^\delta(x_{k+1}^\delta) \leqslant \frac{k}{k+\alpha-1}\Theta^\delta(x_k^\delta) + \frac{\alpha-1}{k+\alpha-1}\Theta^\delta(x_*) + \Delta(\delta)$$
$$+ \frac{(\alpha-1)^2}{2\omega(k+\alpha-1)^2}\left(\left\|w_k^\delta - x_*\right\|^2 - \left\|w_{k+1}^\delta - x_*\right\|^2\right).$$

Equivalently, we can write this as

$$\Theta^\delta(x_{k+1}^\delta) - \Theta^\delta(x_*) \leqslant \frac{k}{k+\alpha-1}\left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*)\right) + \Delta(\delta)$$
$$+ \frac{(\alpha-1)^2}{2\omega(k+\alpha-1)^2}\left(\left\|w_k^\delta - x_*\right\|^2 - \left\|w_{k+1}^\delta - x_*\right\|^2\right).$$

Multiplying by $\frac{2\omega}{\alpha-1}(k+\alpha-1)^2$, we obtain

$$\frac{2\omega}{\alpha-1}(k+\alpha-1)^2\left(\Theta^\delta(x_{k+1}^\delta) - \Theta^\delta(x_*)\right) \leqslant \frac{2\omega}{\alpha-1}k(k+\alpha-1)\left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*)\right)$$
$$+ \frac{2\omega}{\alpha-1}(k+\alpha-1)^2\Delta(\delta) + (\alpha-1)\left(\left\|w_k^\delta - x_*\right\|^2 - \left\|w_{k+1}^\delta - x_*\right\|^2\right),$$

and therefore, since there holds

$$k(k + \alpha - 1) = (k + \alpha - 1)^2 - k(\alpha - 3) - (\alpha - 2)^2 \leqslant (k + \alpha - 1)^2 - k(\alpha - 3),$$

we get that

$$\frac{2\omega}{\alpha - 1}(k + \alpha - 1)^2 \left(\Theta^\delta(x_{k+1}^\delta) - \Theta^\delta(x_*) \right) \leqslant -\frac{2\omega}{\alpha - 1}k(\alpha - 3) \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right)$$
$$+ \frac{2\omega}{\alpha - 1}(k + \alpha - 1)^2 \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) + \frac{2\omega}{\alpha - 1}(k + \alpha - 1)^2 \Delta(\delta)$$
$$+ (\alpha - 1) \left(\left\| w_k^\delta - x_* \right\|^2 - \left\| w_{k+1}^\delta - x_* \right\|^2 \right).$$

Together with the definition (2.8) of $\mathcal{E}^\delta$, this implies

$$\mathcal{E}^\delta(k+1) + \frac{2\omega}{\alpha - 1}k(\alpha - 3) \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) \leqslant \mathcal{E}^\delta(k) + \frac{2\omega}{\alpha - 1}(k + \alpha - 1)^2 \Delta(\delta),$$

or equivalently, after rearranging, we get

$$\mathcal{E}^\delta(k+1) + \frac{2\omega}{\alpha - 1} \left(k(\alpha - 3) \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) - (k + \alpha - 1)^2 \Delta(\delta) \right) \leqslant \mathcal{E}^\delta(k),$$

which concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

Inequality (2.19) is the key ingredient for showing that (1.22), combined with a suitable stopping rule, gives rise to a convergent regularization method. In order to derive a suitable stopping rule, note first that in the case of exact data, i.e. $\delta = 0$, inequality (2.19) reduces to

$$\mathcal{E}^0(k+1) + \frac{2\omega}{\alpha - 1}k(\alpha - 3) \left(\Theta^0(x_k^0) - \Theta^0(x_*) \right) \leqslant \mathcal{E}^0(k). \qquad (2.21)$$

Since by assumption 2.1 the functional Φ^0 is convex, the arguments used in [2] are applicable, and we can deduce the following:

Theorem 2.8. *Let assumption 2.1 hold, let the sequence of iterates x_k^0 and z_k^0 be given by (1.22) with exact data $y = y^\delta$, i.e. $\delta = 0$ and let $\mathcal{S}$ be defined by (2.5). Then the following statements hold:*

- *The sequence $(\mathcal{E}^0(k))$ is non-increasing and $\lim_{k \to \infty} \mathcal{E}^0(k)$ exists.*
- *For each $k \geqslant 0$, there holds*

$$\left\| F(x_k^0) - y \right\|^2 \leqslant \frac{(\alpha - 1)\mathcal{E}^0(0)}{\omega(k + \alpha - 2)^2}, \qquad \left\| w_k^0 - x_* \right\|^2 \leqslant \frac{\mathcal{E}^0(0)}{\alpha - 1}.$$

- *There holds*

$$\sum_{k=1}^\infty k \left\| F(x_k^0) - y \right\|^2 \leqslant \frac{(\alpha - 1)\mathcal{E}^0(1)}{\omega(\alpha - 3)},$$

as well as

$$\sum_{k=1}^\infty k \left\| x_{k+1}^0 - x_k^0 \right\|^2 \leqslant \frac{(\alpha - 1)\mathcal{E}^0(1)}{\omega(\alpha - 3)}.$$

- *There holds*

$$\liminf_{k \to \infty} \left(k^2 \ln(k) \left\| F(x_k^0) - y \right\|^2 \right) = 0,$$

  as well as

$$\liminf_{k \to \infty} \left(k \ln(k) \left\| x_{k+1}^0 - x_k^0 \right\|^2 \right) = 0.$$

- *There exists an $\tilde{x}$ in $\mathcal{S}$, such that the sequence (x_k^0) converges weakly to $\tilde{x}$, i.e.*

$$\lim_{\delta \to 0} \left\langle x_k^0, h \right\rangle = \left\langle \tilde{x}, h \right\rangle, \qquad \forall h \in \mathcal{X}. \tag{2.22}$$

Proof.   The statements follow from facts 1–4, remark 2 and theorem 3 in [2].   □

Thanks to theorem 2.8, we now know that Nesterov's accelerated gradient method (1.22) converges weakly to a solution $\tilde{x}$ from the solution set $\mathcal{S}$ in case of exact data $y = y^\delta$, i.e. $\delta = 0$.

Hence, it remains to consider the behaviour of (1.22) in the case of inexact data y^δ. As mentioned above, the key for doing so is inequality (2.19). We want to use it to show that, similarly to the exact data case, the sequence $(\mathcal{E}^\delta(k))$ is non-increasing up to some $k \in \mathbb{N}$. To do this, note first that $\mathcal{E}^\delta(k)$ is positive as long as

$$\Theta^\delta(x_k^\delta) \geqslant \Theta^\delta(x_*),$$

which is true, as long as

$$\left\| F(x_k^\delta) - y^\delta \right\| \geqslant \delta. \tag{2.23}$$

On the other hand, the term

$$\frac{2\omega}{\alpha - 1} \left(k(\alpha - 3) \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) - (k + \alpha - 1)^2 \Delta(\delta) \right) \tag{2.24}$$

in (2.19) is positive, as long as

$$\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \geqslant \frac{(k + \alpha - 1)^2}{k(\alpha - 3)} \Delta(\delta),$$

which is satisfied, as long as

$$\left\| F(x_k^\delta) - y^\delta \right\|^2 \geqslant \frac{2(k + \alpha - 1)^2}{k(\alpha - 3)} \Delta(\delta) + \delta^2, \tag{2.25}$$

which obviously implies (2.23). These considerations suggest, given a small $\tau > 1$, to choose the stopping index $k_* = k_*(\delta, y^\delta)$ as the smallest integer such that

$$\left\| F(x_{k_*}^\delta) - y^\delta \right\|^2 \leqslant \frac{2(k + \alpha - 1)^2}{k(\alpha - 3)} \Delta(\delta) + \tau^2 \delta^2 < \left\| F(x_k^\delta) - y^\delta \right\|^2, \qquad k_* > k. \tag{2.26}$$

Concerning the well-definedness of k_*, we are able to prove the following

Lemma 2.9.   *Let assumption 2.1 hold, let the sequence of iterates x_k^δ and z_k^δ be given by (1.22) and let c_1 and c_2 be defined by (2.15). Then the stopping index k_* defined by (2.26) with $\tau > 1$ is well-defined and there holds*

$$k_* = \mathcal{O}(\delta^{-1}). \tag{2.27}$$

Proof. By the definition (2.16) of $\Delta(\delta)$ and due to

$$\left\| F(x_k^\delta) - y^\delta \right\|^2 \leqslant \left(\left\| F(x_k^\delta) - y \right\| + \left\| y - y^\delta \right\| \right)^2 \leqslant (2\bar{\omega}\rho + \delta)^2,$$

it follows from (2.26) that for all $k < k_*$ there holds

$$\frac{2(k + \alpha - 1)^2}{k(\alpha - 3)}(c_1\delta + c_2\delta^2) + \tau^2\delta^2 \leqslant (2\bar{\omega}\rho + \delta)^2,$$

which can be rewritten as

$$\frac{(k + \alpha - 1)^2}{k(\alpha - 3)}(c_1\delta + c_2\delta^2) \leqslant 2\bar{\omega}^2\rho^2 + 2\bar{\omega}\rho\delta + (1 - \tau^2)\delta^2 \leqslant 2\bar{\omega}^2\rho^2 + 2\bar{\omega}\rho\delta, \tag{2.28}$$

where we have used that $\tau > 1$. Since the left hand side in the above inequality goes to ∞ for $k \to \infty$, while the right hand side stays bounded, it follows that k_* is finite and hence well-defined for $\delta \neq 0$. Furthermore, since

$$\frac{(k + \alpha - 1)^2}{k(\alpha - 3)} \geqslant \frac{k}{2(\alpha - 3)},$$

which can see by multiplying the above inequality by $k(\alpha - 3)$, and since (2.28) also holds for $k = k_* - 1$, we get

$$\frac{k_* - 1}{2(\alpha - 3)}(c_1\delta + c_2\delta^2) \leqslant 2\bar{\omega}^2\rho^2 + 2\bar{\omega}\rho\delta.$$

Reordering the terms, we arrive at

$$k_* \leqslant 2(\alpha - 3)\left(\frac{2\bar{\omega}^2\rho^2 + 2\bar{\omega}\rho\,\delta}{c_1\delta + c_2\delta^2} \right) + 1$$

from which the assertion now immediately follows.   □

    The rate $k_* = \mathcal{O}(\delta^{-1})$ given in (2.27) for the iteration method (1.22) should be compared with the corresponding result [23, corollary 2.3] for Landweber iteration (1.4), where one only obtains $k_* = \mathcal{O}(\delta^{-2})$. In order to obtain the rate $k_* = \mathcal{O}(\delta^{-1})$ for Landweber iteration, apart from others, a source condition of the form

$$x^\dagger - x_0 \in \mathcal{R}(F'(x^\dagger)^*) \tag{2.29}$$

has to hold, which is not required for Nesterov's accelerated gradient method (1.22).

    Before we turn to our main result, we first prove a couple of important consequences of (2.19) and the stopping rule (2.26).

Proposition 2.10. *Let assumption 2.1 be satisfied, let x_k^δ and z_k^δ be defined by (1.22) and let $\mathcal{E}^\delta$ be defined by (2.8). Assuming that the stopping index k_* is determined by (2.26) with some $\tau > 1$, then, for all $0 \leqslant k \leqslant k_*$, the sequence $(\mathcal{E}^\delta(k))$ is non-increasing and in particular, $\mathcal{E}^\delta(k) \leqslant \mathcal{E}^\delta(0)$. Furthermore, for all $0 \leqslant k \leqslant k_*$ there holds*

$$\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \leqslant \frac{(\alpha - 1)\mathcal{E}^\delta(0)}{2\omega(k + \alpha - 2)^2}, \tag{2.30}$$

as well as

$$\left\| w_k^\delta - x_* \right\|^2 \leqslant \frac{\mathcal{E}^\delta(0)}{(\alpha - 1)}, \tag{2.31}$$

and

$$\sum_{k=1}^{k_*-1} \left(k \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) - \frac{(k + \alpha - 1)^2}{(\alpha - 3)} \Delta(\delta) \right) \leqslant \frac{(\alpha - 1)\mathcal{E}^\delta(1)}{2\omega(\alpha - 3)}. \tag{2.32}$$

Proof. Due to the definition of the stopping rule (2.26) and the arguments preceding it, the term (2.24) is positive for all $k \leqslant k_* - 1$. Hence, due to (2.19), $\mathcal{E}^\delta(k)$ is non-increasing for all $k \leqslant k_*$ and in particular, $\mathcal{E}^\delta(k) \leqslant \mathcal{E}^\delta(0)$. From this observation, (2.30) and (2.31) immediately follow from the definition (2.8) of $\mathcal{E}^\delta(k)$.

Furthermore, rearranging (2.19) we have

$$\frac{2\omega(\alpha - 3)}{\alpha - 1} \left(k \left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*) \right) - \frac{(k + \alpha - 1)^2}{(\alpha - 3)} \Delta(\delta) \right) \leqslant \mathcal{E}^\delta(k) - \mathcal{E}^\delta(k + 1).$$

Now, summing over this inequality and using telescoping and the fact that $\mathcal{E}^\delta(k_*) \geqslant 0$ we immediately arrive at (2.32), which concludes the proof. □

From the above proposition, we are able to deduce two interesting corollaries.

Corollary 2.11. *Under the assumptions of proposition 2.10 there holds*

$$\left\| F(x_k^\delta) - y^\delta \right\|^2 \leqslant \frac{2(\alpha - 1)\mathcal{E}^\delta(0)}{\omega(k + \alpha - 2)^2} + \delta^2, \qquad 0 \leqslant k \leqslant k_*. \tag{2.33}$$

Proof. Using the fact that both $x_k^\delta, x_* \in \mathcal{B}_{2\rho}(x_0)$, it follows from the definition of Θ^δ that $\Theta^\delta(x_k^\delta) = \Phi^\delta(x_k^\delta)$ and $\Theta^\delta(x_*) = \Phi^\delta(x_*)$. Hence, inequality (2.30) yields

$$\left\| F(x_k^\delta) - y^\delta \right\|^2 \leqslant \frac{2(\alpha - 1)\mathcal{E}^\delta(0)}{\omega(k + \alpha - 2)^2} + \left\| y - y^\delta \right\|^2, \qquad 0 \leqslant k \leqslant k_*,$$

from which, using $\left\| y - y^\delta \right\| \leqslant \delta$, the statement immediately follows. □

Corollary 2.12. *Under the assumptions of proposition 2.10 there holds*

$$k_*(k_* - 1) \leqslant \left(\frac{2(\alpha - 1)\mathcal{E}^\delta(1)}{\omega(\alpha - 3)(\tau^2 - 1)} \right) \frac{1}{\delta^2}.$$

Proof. Using the fact that both $x_k^\delta, x_* \in \mathcal{B}_{2\rho}(x_0)$, it follows from the definition of Θ^δ that $\Theta^\delta(x_k^\delta) = \Phi^\delta(x_k^\delta)$ and $\Theta^\delta(x_*) = \Phi^\delta(x_*)$ Hence, it follows with $\left\| y - y^\delta \right\| \leqslant \delta$ that

$$k\left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*)\right) - \frac{(k+\alpha-1)^2}{(\alpha-3)}\Delta(\delta)$$
$$\geqslant \frac{k}{2}\left(\left\|F(x_k^\delta) - y^\delta\right\|^2 - \delta^2\right) - \frac{(k+\alpha-1)^2}{(\alpha-3)}\Delta(\delta).$$

Together with the definition of the stopping rule (2.26), this implies that for all $k \leqslant k_* - 1$

$$k\left(\Theta^\delta(x_k^\delta) - \Theta^\delta(x_*)\right) - \frac{(k+\alpha-1)^2}{(\alpha-3)}\Delta(\delta) > k\frac{(\tau^2-1)\delta^2}{2}.$$

Using this in (2.32) yields

$$\frac{(\tau^2-1)\delta^2}{2}\sum_{k=1}^{k_*-1}k \leqslant \frac{(\alpha-1)\mathcal{E}^\delta(1)}{2\omega(\alpha-3)},$$

from which the statement now immediately follows. $\qquad\square$

Again, this shows that $k_* = \mathcal{O}(\delta^{-1})$, i.e. $k_* \leqslant c\delta^{-1}$, however this time the constant c does not depend on c_1 and c_2, an observation which we use when analysing (1.22) under slightly different assumptions than assumption 2.1 below.

We are now able to prove one of our main results:

Theorem 2.13. *Let assumption 2.1 hold and let the iterates x_k^δ and z_k^δ be defined by (1.22). Furthermore, let $k_* = k_*(\delta, y^\delta)$ be determined by (2.26) with some $\tau > 1$ and let the solution set $\mathcal{S}$ be given by (2.5). Then there exists an $\tilde{x} \in \mathcal{S}$ and a subsequence $\tilde{x}_{k_*}^\delta$ of $x_{k_*}^\delta$ which converges weakly to $\tilde{x}$ as $\delta \to 0$. If $\mathcal{S}$ is a singleton, then $x_{k_*}^\delta$ converges weakly to the then unique solution $\tilde{x} \in \mathcal{S}$.*

Proof. This proof follows some ideas of [15]. Let $y_n := y^{\delta_n}$ be a sequence of noisy data satisfying $\|y - y_n\| \leqslant \delta_n$. Furthermore, let $k_n := k_*(\delta_n, y_n)$ be the stopping index determined by (2.26) applied to the pair (δ_n, y_n). There are two cases. First, assume that k is a finite accumulation point of k_n. Without loss of generality, we can assume that $k_n = k$ for all $n \in \mathbb{N}$. Thus, from (2.26), it follows that

$$\left\|F(x_k^{\delta_n}) - y_n\right\| < \frac{2(k+\alpha-1)^2}{k(\alpha-3)}\Delta(\delta_n) + \tau^2\delta_n^2,$$

which, together with the triangle inequality, implies

$$\left\|F(x_k^{\delta_n}) - y\right\| \leqslant \left\|F(x_k^{\delta_n}) - y_n\right\| + \|y_n - y\| \leqslant \frac{2(k+\alpha-1)^2}{k(\alpha-3)}\Delta(\delta_n) + \tau^2\delta_n^2 + \delta_n.$$

Since for fixed k the iterates x_k^δ depend continuously on the data y^δ, by taking the limit $n \to \infty$ in the above inequality we can derive

$$x_k^{\delta_n} \to x_k^0, \qquad F(x_k^{\delta_n}) \to F(x_k^0) = y, \quad \text{as} \quad n \to \infty.$$

For the second case, assume that $k_n \to \infty$ as $n \to \infty$. Since $x_{k_n}^{\delta_n} \in \mathcal{B}_{2\rho}(x_0)$, it is bounded and hence, has a weakly convergent subsequence $x_{\tilde{k}_n}^{\tilde{\delta}_n}$, corresponding to a subsequence $\tilde{\delta}_n$ of δ_n

and $\tilde{k}_n := k_*(\tilde{\delta}_n, y^{\delta_n})$. Denoting the weak limit of $x_{\tilde{k}_n}^{\tilde{\delta}_n}$ by $\tilde{x}$, it remains to show that $\tilde{x} \in \mathcal{S}$. For this, observe that it follows from (2.33) that

$$\left\| F(x_{\tilde{k}_n}^{\tilde{\delta}_n}) - y^\delta \right\|^2 \leqslant \frac{2(\alpha - 1)\mathcal{E}^{\tilde{\delta}_n}(0)}{\omega(\tilde{k}_n + \alpha - 2)^2} + \tilde{\delta}_n^2 \longrightarrow 0, \quad \text{as} \quad n \to \infty$$

where we have used that $\tilde{k}_n \to \infty$ and $\tilde{\delta}_n \to 0$ as $n \to \infty$, which follows from the assumption that so do the sequences k_n and δ_n, and the fact that $\mathcal{E}^\delta(0)$ stays bounded for $\delta \to 0$. Hence, since we know that $y^\delta \to y$ as $\delta \to 0$, we can deduce that

$$F\left(x_{\tilde{k}_n}^{\tilde{\delta}_n}\right) \to y, \quad \text{as} \quad n \to \infty,$$

and therefore, using the weak sequential closedness of F on $\mathcal{B}_{2\rho}(x_0)$, we deduce that $F(\tilde{x}) = y$, i.e. $\tilde{x} \in \mathcal{S}$, which was what we wanted to show.

It remains to show that if $\mathcal{S}$ is a singleton then $x_{k_*}^\delta$ converges weakly to $\tilde{x}$. Since this was already proven above in the case that k_n has a finite accumulation point, it remains to consider the second case, i.e. $k_n \to \infty$. For this, consider an arbitrary subsequence of $x_{k_*}^\delta$. Since this sequence is bounded, it has a weakly convergent subsequence which, by the same arguments as above, converges to a solution $\tilde{x} \in \mathcal{S}$. However, since we have assumed that $\mathcal{S}$ is a singleton, it follows that $x_{k_*}^\delta$ converges weakly to $\tilde{x}$, which concludes the proof. $\qquad\square$

Remark.   In theorem 2.13, we have shown weak subsequential convergence to an element $\tilde{x}$ in the solution set $\mathcal{S}$. However, this element might be different from the x_0-minimum norm solution $x^\dagger$ defined by (2.6), unless of course in case that $\mathcal{S}$ is a singleton.

Furthermore, it follows from the proof of theorem 2.13 that any (weakly) convergent subsequence of $x_{k_*}^\delta$ also converges (weakly) to an element in the solution set $\mathcal{S}$.

3. Convergence analysis II

Some simplifications of the above presented convergence analysis are possible if we assume that instead of only Φ^0, all the functionals Φ^δ are convex. Hence, for the remainder of this section, we work with the following:

Assumption 3.1.   *Let ρ be a positive number such that $\mathcal{B}_{6\rho}(x_0) \subset \mathcal{D}(F)$.*

1. *The operator $F : \mathcal{D}(F) \subset \mathcal{X} \to \mathcal{Y}$ is continuously Fréchet differentiable between the real Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with inner products $\langle \,.,. \,\rangle$ and norms $\|.\|$. Furthermore, let F be weakly sequentially closed on $\mathcal{B}_{2\rho}(x_0)$.*
2. *The equation $F(x) = y$ has a solution $x_* \in \mathcal{B}_\rho(x_0)$.*
3. *The data y^δ satisfies $\|y - y^\delta\| \leqslant \delta$.*
4. *The functionals Φ^δ are convex and have Lipschitz continuous gradients $\nabla\Phi^\delta$ with uniform Lipschitz constant L on $\mathcal{B}_{6\rho}(x_0)$, i.e.*

$$\Phi^\delta(\lambda x_1 + (1 - \lambda)x_2) \leqslant \lambda\Phi^\delta(x_1) + (1 - \lambda)\Phi^\delta(x_2), \qquad \forall x_1, x_2 \in \mathcal{B}_{6\rho}(x_0), \tag{3.1}$$

$$\left\|\nabla\Phi^\delta(x_1) - \nabla\Phi^\delta(x_2)\right\| \leqslant L \left\|x_1 - x_2\right\|, \qquad \forall x_1, x_2 \in \mathcal{B}_{6\rho}(x_0).$$

5. *For α in (1.22) there holds $\alpha > 3$ and the scaling parameter ω satisfies $0 < \omega < \frac{1}{L}$.*

Note that assumption 3.1 is only a special case of assumption 2.1. Hence, the above convergence analysis presented above is applicable and we get weak convergence of the iterates of (1.22). However, the stopping rule (2.26) depends on the constants c_1 and c_2 defined by (2.15), which are not always available in practise. Fortunately, using the assumption 3.1, we can get rid of c_1 and c_2. The key idea is to observe that the following lemma holds:

Lemma 3.1. *Under assumption 3.1, for all $x, z \in \mathcal{B}_{6\rho}(x_0)$ there holds*

$$\Theta^\delta(z - \omega G_\omega^\delta(z)) \leqslant \Theta^\delta(x) + \left\langle G_\omega^\delta(z), z - x \right\rangle - \frac{\omega}{2}\left\|G_\omega^\delta(z)\right\|^2.$$

Proof. This follows from the convexity of Θ^δ in the same way as in lemma 2.2. $\qquad\square$

From the above lemma, it follows that the results of corollary 2.5 and proposition 2.6 hold with $\Delta(\delta) = 0$. Therefore, the stopping rule (2.26) simplifies to

$$\left\|F(x_{k_*}^\delta) - y^\delta\right\| \leqslant \tau\delta < \left\|F(x_k^\delta) - y^\delta\right\|, \qquad k_* \geqslant k, \tag{3.2}$$

for some $\tau > 1$, which is nothing else than the discrepancy principle (1.6). Note that in contrast to (2.26), only the noise level δ needs to be known in order to determine the stopping index k_*. With the same arguments as above, we are now able to prove our second main result:

Theorem 3.2. *Let assumption 3.1 hold and let the iterates x_k^δ and z_k^δ be defined by (1.22). Furthermore, let $k_* = k_*(\delta, y^\delta)$ be determined by (3.2) with some $\tau > 1$ and let the solution set $\mathcal{S}$ be given by (2.5). Then for the stopping index k_* there holds $k_* = \mathcal{O}(\delta^{-1})$. Furthermore, there exists an $\tilde{x} \in \mathcal{S}$ and a subsequence $\tilde{x}_{k_*}^\delta$ of $x_{k_*}^\delta$ which converges weakly to $\tilde{x}$ as $\delta \to 0$. If $\mathcal{S}$ is a singleton, then $x_{k_*}^\delta$ converges weakly to the then unique solution $\tilde{x} \in \mathcal{S}$.*

Proof. The proof of this theorem is analogous to the proof of theorem 2.13. The only main difference is the well definedness of k_*, which now cannot be derived from lemma 2.9 but follows from (2.32) by corollary 2.12, which also yields $k_* = \mathcal{O}(\delta^{-1})$. $\qquad\square$

Remark. As in the remark after theorem 2.13, a short inspection of the proof of theorem 3.2 shows again that also in this case any (weakly) convergent subsequence of $x_{k_*}^\delta$ also converges (weakly) to an element in the solution set $\mathcal{S}$.

Remark. Note that since theorem 3.2 only gives an asymptotic result, i.e. for $\delta \to 0$, the requirement in assumption 3.1 that the functionals Φ^δ have to be convex for all $\delta > 0$ can be relaxed to $0 \leqslant \delta \leqslant \bar{\delta}$, as long as we only consider data y^δ satisfying the noise constraint $\left\|y - y^\delta\right\| \leqslant \delta \leqslant \bar{\delta}$.

Remark. Note that if the functionals Φ^δ are globally convex and uniformly Lipschitz continuous, which is for example the case if F is a bounded linear operator, then one can choose ρ arbitrarily large in the definition of Ψ. Now, as we have seen above, the proximal mapping $\text{prox}_{\omega\Psi}(.)$ is nothing else than the projection onto $\mathcal{B}_{2\rho}(x_0)$. This implies that for practical purposes, $\text{prox}_{\omega\Psi}(.)$ may be dropped in (1.22), which means that one effectively uses (1.18) instead of (1.22).

4. Strong convexity and nonlinearity conditions

In this section, we consider the question of strong convergence of the iterates of (1.22) and comment on the connection between the assumption of local convexity and the (weak) tangential cone condition.

Concerning the strong convergence of the iterates of (1.22) and (1.18), note that it could be achieved if the functional Φ^0 were locally strongly convex, i.e. if

$$\langle F'(x_1)^*(F(x_1) - y) - F'(x_2)^*(F(x_2) - y), x_1 - x_2 \rangle \geqslant \alpha \|x_1 - x_2\|^2, \qquad (4.1)$$
$$\forall x_1, x_2 \in \mathcal{B}_{2\rho}(x_0),$$

since then, for the choice of $x_1 = x_k^0$ and $x_2 = x_*$, one gets

$$\alpha \left\| x_k^0 - x_* \right\| \leqslant \langle F'(x_k^0)^*(F(x_k^0) - y), x_k^0 - x_* \rangle \leqslant 2\bar{\omega}\rho \left\| F(x_k^0) - y \right\|,$$

from which, since we have $\left\| F(x_k^0) - y \right\| \to 0$ as $\delta \to 0$, it follows that x_k^δ converges strongly to x_* as $\delta \to 0$. Hence, retracing the proof of theorem 2.13, one would get

$$\lim_{\delta \to 0} x_{k_*}^\delta = x_*.$$

Unfortunately, already for linear ill-posed operators $F = A$, strong convexity of the form (4.1) cannot be satisfied, since then one would get

$$\|Ax_1 - Ax_2\| \geqslant \alpha \|x_1 - x_2\|, \qquad \forall x_1, x_2 \in \mathcal{B}_{2\rho}(x_0),$$

which already implies the well-posedness of $Ax = y$ in $\mathcal{B}_{2\rho}(x_0)$. However, defining

$$\mathcal{M}_\tau(A) := \left\{ x \in \mathcal{B}_{2\rho} \,|\, \exists w \in \mathcal{Y}, \|w\| \leqslant \tau, \, x - x^\dagger = A^*w \right\}, \qquad (4.2)$$

it was shown in [16, lemma 3.3] that there holds

$$\left\| x - x^\dagger \right\|^2 \leqslant \tau \left\| Ax - Ax^\dagger \right\|, \qquad \forall x \in \mathcal{M}_\tau(A).$$

Hence, if one could show that $x_k^0 \in \mathcal{M}_\tau$ for some $\tau > 0$ and all $k \in \mathbb{N}$, then it would follow that

$$\left\| x_k^0 - x^\dagger \right\|^2 \leqslant \tau \left\| Ax_k^0 - y \right\|, \qquad \forall x \in \mathcal{M}_\tau(A),$$

from which strong convergence of x_k^0, and consequently also of $x_{k_*}^\delta$ to $x^\dagger$ would follow. In essence, this was done in [28] with tools from spectral theory in the classical framework for analysing linear ill-posed problem [9] under the source condition $x^\dagger \in \mathcal{R}(A^*)$.

Remark. Note that it is sometimes possible, given weak convergence of a sequence $x_k \in \mathcal{X}$ to some element $\tilde{x} \in \mathcal{X}$, to infer strong convergence of x_k to $\tilde{x}$ in a weaker topology. For example, if $x_k \in H^1(0,1)$ converges weakly to $\tilde{x}$ in the $H^1(0,1)$ norm, then it follows that x_k converges strongly to $\tilde{x}$ with respect to the $L^2(0,1)$ norm. Many generalizations of this example are possible. Note further that in finite dimensions, weak and strong convergence coincide.

In the remaining part of this section, we want to comment on the connection of the local convexity assumption (2.1) to other nonlinearity conditions like (1.7) and (1.8) commonly used in the analysis of nonlinear-inverse problems.

First of all, note that due to the results of Kindermann [24], we know that both convexity and the (weak) tangential cone condition imply weak convergence of Landweber iteration (1.4). However, it is not entirely clear in which way those conditions are connected.

One connection of the two conditions was given in [32], where it was shown that the nonlinearity condition implies a certain directional convexity condition. Another connection was provided in [24], where it was shown that the tangential cone condition implies a quasi-convexity condition. However, it is not clear whether or not the tangential cone condition implies convexity or not. What we can say is that convexity does not imply the (weak) tangential cone condition, which is shown in the following

Example 4.1. Consider the operator $F : H^1(0, 1) \to L^2(0, 1)$ defined by

$$F(x)(s) := \int_0^s x(t)^2 \, \mathrm{d}t. \tag{4.3}$$

This nonlinear Hammerstein operator was extensively treated as an example problem for nonlinear inverse problems (see for example [15, 27]). It is well known that for this operator the tangential cone condition is satisfied around $x^\dagger$ as long as $x^\dagger \geqslant c > 0$. However, the (weak) tangential cone condition is not satisfied in case that $x^\dagger \equiv 0$. Moreover, it can easily be seen (for example from (5.1)) that $\Phi^0(x)$ is globally convex, which shows that convexity does not imply the tangential cone condition.

5. Example problems

In this section, we consider two examples to which we apply the theory developed above. Most importantly, we prove the local convexity assumption for both Φ^0 and Φ^δ, with δ small enough. Furthermore, based on these example problems, we present some numerical results, demonstrating the usefulness of method (1.22), and supporting the findings of [17–19, 21, 28], which are also shortly discussed.

For this, note that if F is twice continuously Fréchet differentiable, then convexity of Φ^δ is equivalent to positive semi-definiteness of its second Fréchet derivative [31]. More precisely, we have that (3.1) is equivalent to

$$\|F'(x)h\|^2 + \left\langle F(x) - y^\delta, F''(x)(h, h) \right\rangle \geqslant 0, \qquad \forall x \in \mathcal{B}_{6\rho}(x_0), \, \forall h \in \mathcal{D}(F), \tag{5.1}$$

which is our main tool for the upcoming analysis.

5.1. Example 1—Nonlinear diagonal operator

For our first (academic) example, we look at the following class of nonlinear diagonal operators

$$F : \ell^2 \to \ell^2, \qquad x := (x_{(n)})_{n \in \mathbb{N}} \mapsto \sum_{n=1}^\infty f_n(x_{(n)}) \, e_{(n)}$$

where $\left(e_{(n)}\right)_{n \in \mathbb{N}}$ is the canonical orthonormal basis of ℓ^2. These operators are reminiscent of the singular value decomposition of compact linear operators. Here we consider the special choice

$$f_n(z) := \frac{1}{n} \cdot \begin{cases} z^2, & n \leqslant M, \\ z, & n > M, \end{cases} \tag{5.2}$$

for some fixed integer $M > 0$. For this choice, F takes the form

$$F(x) = \sum_{n=1}^{M} \frac{1}{n} x_{(n)}^2 e_{(n)} + \sum_{n=M+1}^{\infty} \frac{1}{n} x_{(n)} e_{(n)}.$$

It is easy to see that F is a well-defined, twice continuously Fréchet differentiable operator with

$$F'(x)h = 2 \sum_{n=1}^{M} \frac{1}{n} x_{(n)} h_{(n)} e_{(n)} + 2 \sum_{n=M+1}^{\infty} \frac{1}{n} h_{(n)} e_{(n)},$$

$$F''(x)(h, w) = 2 \sum_{n=1}^{M} \frac{1}{n} h_{(n)} w_{(n)} e_{(n)}.$$

Furthermore, note that solving $F(x) = y$ is equivalent to

$$x_{(n)} = n \begin{cases} \sqrt{y_{(n)}}, & n \leqslant M, \\ y_{(n)}, & n > M, \end{cases}$$

from which it is easy to see that we are dealing with an ill-posed problem.

We now turn to the convexity of $\Phi^\delta(x)$ around a solution $x^\dagger$.

Proposition 5.1. *Let $x^\dagger$ be a solution of $F(x) = y$ such that $\left| x_{(n)}^\dagger \right| > 0$ holds for all $n \in \{1, \dots, M\}$. Furthermore, let $\rho > 0$ and $\bar\delta \geqslant 0$ be small enough such that*

$$(x_{(n)}^\dagger)^2 \geqslant 28 |x_{(n)}^\dagger| \rho + \bar\delta \left(2 \|y\|_{\ell^2} + \bar\delta \right), \qquad \forall n \in \{1, \dots, M\}, \tag{5.3}$$

and let $x_0 \in \mathcal{B}_\rho(x^\dagger)$. Then for all $0 \leqslant \delta \leqslant \bar\delta$, the functional $\Phi^\delta(x)$ is convex in $\mathcal{B}_{6\rho}(x_0)$.

Proof. Due to (5.1) it is sufficient to show that

$$0 \leqslant \|F'(x)h\|^2 + \left\langle F(x) - y^\delta, F''(x)(h, h) \right\rangle =$$
$$= \|F'(x)h\|^2 + \left\langle F(x) - y, F''(x)(h, h) \right\rangle + \left\langle y^\delta - y, F''(x)(h, h) \right\rangle.$$

Using the definition of F, the fact that e_n is an orthonormal basis of ℓ^2 and that $F(x^\dagger) = y$, this inequality can be rewritten into

$$2 \left(\sum_{n=1}^{M} \frac{1}{n^2} x_{(n)}^2 h_{(n)}^2 + \sum_{n=M+1}^{\infty} \frac{1}{n^2} h_{(n)}^2 \right) + 2 \sum_{n=1}^{M} (x_{(n)}^2 - (x_{(n)}^\dagger)^2) h_{(n)}^2$$

$$+ 2 \sum_{n=1}^{M} (y_{(n)}^2 - (y_{(n)}^\delta)^2) h_{(n)}^2 \geqslant 0,$$

which, after simplification, becomes

$$2 \sum_{n=1}^{M} h_{(n)}^2 \left(2x_{(n)}^2 - (x_{(n)}^\dagger)^2 + y_{(n)}^2 - (y_{(n)}^\delta)^2 \right) + 2 \sum_{n=M+1}^{\infty} \frac{1}{n^2} h_{(n)}^2 \geqslant 0.$$

Since the right of the above two sums is always positive, in order for the above inequality to be satisfied it suffices to show that

$$2x_{(n)}^2 - (x_{(n)}^\dagger)^2 + y_{(n)}^2 - (y_{(n)}^\delta)^2 \geqslant 0, \qquad \forall n \in \{1, \dots, M\}. \tag{5.4}$$

Now, since by the triangle inequality we have

$$
\left| y_{(n)}^2 - (y_{(n)}^\delta)^2 \right| = \left| y_{(n)} - y_{(n)}^\delta \right| \left| y_{(n)} + y_{(n)}^\delta \right| \leqslant \left\| y - y^\delta \right\|_{\ell^2} \left\| y + y^\delta \right\|_{\ell^2}
$$
$$
\leqslant \delta \left(2 \left\| y \right\|_{\ell^2} + \left\| y - y^\delta \right\|_{\ell^2} \right) \leqslant \delta \left(2 \left\| y \right\|_{\ell^2} + \delta \right),
$$

(5.5)

it follows that in order to prove (5.4) it suffices to show

$$
2x_{(n)}^2 - (x_{(n)}^\dagger)^2 - \delta \left(2 \left\| y \right\|_{\ell^2} + \delta \right) \geqslant 0, \qquad \forall n \in \{1, \dots, M\}.
$$

Now, writing $x = x^\dagger + \varepsilon$, this can be rewritten into

$$
(x_{(n)}^\dagger)^2 + 4x_{(n)}^\dagger \varepsilon_{(n)} + 2\varepsilon_{(n)}^2 - \delta \left(2 \left\| y \right\|_{\ell^2} + \delta \right) \geqslant 0, \qquad \forall n \in \{1, \dots, M\}.
$$

Since $\varepsilon_{(n)}^2 \geqslant 0$, the above inequality is satisfied given that

$$
(x_{(n)}^\dagger)^2 - 4 \left| x_{(n)}^\dagger \right| \left| \varepsilon_{(n)} \right| - \delta \left(2 \left\| y \right\|_{\ell^2} + \delta \right) \geqslant 0, \qquad \forall n \in \{1, \dots, M\}.
$$

However, since $\left| \varepsilon_{(n)} \right| \leqslant \left\| \varepsilon \right\|_{\ell^2} = \left\| x - x^\dagger \right\|_{\ell^2} \leqslant \left\| x - x_0 \right\|_{\ell^2} + \left\| x_0 - x^\dagger \right\|_{\ell^2} \leqslant 7\rho$, this follows immediately from (5.3), which concludes the proof. $\qquad \square$

Remark. Due to $\left| x_{(n)}^\dagger \right| \leqslant \left\| x^\dagger \right\|_{\ell^2}$, condition (5.3) is satisfied given that

$$
\min_{n=1,\dots,M} \left\{ (x_{(n)}^\dagger)^2 \right\} \geqslant 28 \left\| x^\dagger \right\|_{\ell^2} \rho + \bar{\delta} \left(\left\| y \right\|_{\ell^2} + \bar{\delta} \right),
$$

which can always be satisfied given that $\left| x_{(n)}^\dagger \right| > 0$ for all $n \in \{1, \dots, M\}$.

After proving local convexity of the residual functional around the solution, we now proceed to demonstrate the usefulness of (1.22) based on the following numerical

Example 5.1. For this example we choose f_n as in (5.2) with $M = 100$. For the exact solution $x^\dagger$ we take the sequence $x_{(n)}^\dagger = 100/n$ which leads to the exact data

$$
y_{(n)} = F(x^\dagger)_{(n)} = \begin{cases} 10^4/n^3, & n \leqslant 100, \\ 10^2/n^2, & n > 100. \end{cases}
$$

Hence, condition (5.4) reads as follows

$$
10^4/n^2 \geqslant 28(10^2/n)\rho + \bar{\delta}(2 \left\| y \right\|_{\ell^2} + \bar{\delta}), \qquad \forall n \in \{1, \dots, 100\}.
$$

Therefore, the functional Φ^0 is convex in $\mathcal{B}_{6\rho}(x_0)$ given that $\rho \leqslant 1/28 \approx 0.036$, which is for example the case for the choice

$$
x_0 = x^\dagger + \left((-1)^n \frac{\rho\sqrt{6}}{\pi n} \right)_{n \in \mathbb{N}}.
$$

(5.6)

Furthermore, for any noise level $\bar{\delta}$ small enough, one has that for all $\delta \leqslant \bar{\delta}$ the functional Φ^δ is convex in $\mathcal{B}_{6\rho}(x_0)$ as long as

$$\rho \leqslant \frac{10^4/n - n\,\bar{\delta}(2\,\|y\|_{\ell^2} + \bar{\delta})}{2800}, \qquad \forall\, n \in \{1, \dots, 100\},$$

which for example is satisfied if

$$\rho \leqslant \frac{1 - \bar{\delta}(2\,\|y\|_{\ell^2} + \bar{\delta})}{28}.$$

For numerically treating the problem, instead of considering full sequences $x = \big(x_{(n)}\big)_{n\in\mathbb{N}}$, we only consider $\vec{x} = \big(x_{(n)}\big)_{n=1,\dots,N}$ where we choose $N = 200$ in this example. This means that we are considering the following discretized version of F:

$$F_n(\vec{x}) = \sum_{n=1}^{100} \frac{1}{n} x_{(n)}^2 e_{(n)} + \sum_{n=101}^{200} \frac{1}{n} x_{(n)} e_{(n)}.$$

We now compare the behaviour of method (1.22) with its non-accelerated Landweber counterpart (1.4) when applied to the problem with $x^\dagger$ and x_0 as defined above. For both methods, we choose the same scaling parameter $\omega = 3.2682 \times 10^{-5}$ estimated from the norm of $F(x^\dagger)$ and we stop the iteration with the discrepancy principle (1.6) with $\tau = 1$. Furthermore, random noise with a relative noise level of 0.001% was added to the data to arrive at the noisy data y^δ and, following the argument presented after (3.2) and since the iterates x_k^δ remain bounded even without it, we drop the proximal operator $\mathrm{prox}_{\omega\Psi}(.)$ in (1.22). The results of the experiments, computed in MATLAB, are displayed in table 1. The speedup both in time and in the number of iterations achieved by Nesterov's acceleration scheme is obvious. Not only does (1.22) satisfy the discrepancy principle much earlier than (1.4), but also the relative error is even a bit smaller for method (1.22).

5.2. Example 2—Auto-convolution operator

Next we look at an example involving an auto-convolution operator. Due to its importance in laser optics, the auto-convolution problem has been extensively studied in the literature [1, 5, 11], its ill-posedness has been shown in [7, 10, 12] and its special structure was successfully exploited in [29]. For our purposes, we consider the following version of the auto-convolution operator

$$F : L^2(0,1) \to L^2(0,1), \qquad F(x)(s) := (x * x)(s) := \int_0^1 x(s-t)x(t)\, \mathrm{d}t, \tag{5.7}$$

where we interpret functions in $L^2(0,1)$ as 1-periodic functions on $\mathbb{R}$. For the following, denote by $(e_{(n)})_{n\in\mathbb{Z}}$ the canonical real Fourier basis of $L^2(0,1)$, i.e.

$$e_{(n)}(t) := \begin{cases} 1, & n = 0, \\ \sqrt{2}\sin(2\pi nt), & n \geqslant 1, \\ \sqrt{2}\cos(2\pi nt), & n \leqslant -1, \end{cases} \qquad t \in (0,1),$$

and by $x_{(n)} := \big\langle x, e_{(n)} \big\rangle$ the Fourier coefficients of x. It follows that

$$x * w = \sum_{n\in\mathbb{Z}} x_{(n)} w_{(n)} e_{(n)}. \tag{5.8}$$

Table 1. Comparison of Landweber iteration (1.4) and its Nesterov accelerated version (1.22) when applied to the diagonal operator problem considered in example 5.1.

Method	k_*	Time (s)	$\left\|x^\dagger - x_k^\delta\right\| / \left\|x^\dagger\right\|$ (%)
Landweber	82	0.057	0.0109
Nesterov	23	0.019	0.0108

It was shown in [8] that if only finitely many Fourier components $x_{(n)}$ are non-zero, then a variational source condition is satisfied leading to convergence rates for Tikhonov regularization. We now use this assumption of a sparse Fourier representation to prove convexity of Φ^δ for the auto-convolution operator in the following

Proposition 5.2. *Let $x^\dagger$ be a solution of $F(x) = y$ such that there exists an index set $\Lambda_N \subset \mathbb{Z}$ with $|\Lambda_N| = N$ such that for the Fourier coefficients $x_{(n)}^\dagger$ of $x^\dagger$ there holds*

$$x_{(n)}^\dagger = 0, \qquad \forall n \in \mathbb{Z} \setminus \Lambda_N.$$

Furthermore, let $\rho > 0$ and $\bar{\delta} \geqslant 0$ be small enough such that

$$(x_{(n)}^\dagger)^2 \geqslant 28|x_{(n)}^\dagger|\,\rho + \bar{\delta}\left(2\,\|y\|_{L^2} + \bar{\delta}\right), \qquad \forall k \in \Lambda_N \tag{5.9}$$

and let $x_0 \in \mathcal{B}_\rho(x^\dagger)$. Then for all $0 \leqslant \delta \leqslant \bar{\delta}$, the functional $\Phi^\delta(x)$ is convex in $\mathcal{B}_{6\rho}(x_0)$.

Proof. As in the previous example, we want to show that (5.1) is satisfied, which, due to (5.8) and the fact that the $e_{(n)}$ form an orthonormal basis is equivalent to

$$\sum_{n \in \mathbb{Z}} x_{(n)}^2 h_{(n)}^2 + \sum_{n \in \mathbb{Z}} (x_{(n)}^2 - (x_{(n)}^\dagger)^2) h_{(n)}^2 + \sum_{n \in \mathbb{Z}} ((y_{(n)}^\delta)^2 - y_{(n)}^2) h_{(n)}^2 \geqslant 0,$$

which, after simplification, becomes

$$\sum_{n \in \mathbb{Z}} h_{(n)}^2 \left(2x_{(n)}^2 - (x_{(n)}^\dagger)^2 + (y_{(n)}^\delta)^2 - y_{(n)}^2\right) \geqslant 0,$$

and hence, it is sufficient to show that

$$2x_{(n)}^2 - (x_{(n)}^\dagger)^2 + (y_{(n)}^\delta)^2 - y_{(n)}^2 \geqslant 0, \qquad \forall n \in \mathbb{Z}. \tag{5.10}$$

Note that this is essentially the same condition as (5.4) in the previous example, apart from that here we have to show the inequality for all $n \in \mathbb{Z}$. However, if $n \notin \Lambda_N$, then $x_{(n)}^\dagger = y_{(n)} = 0$ and hence, (5.10) is trivially satisfied. Hence, it remains to prove (5.10) only for $n \in \Lambda_N$. For this, we write $x_{(n)} = x_{(n)}^\dagger + \varepsilon_{(n)}$, which allows us to rewrite (5.4) into

$$(x_{(n)}^\dagger)^2 + 4x_{(n)}^\dagger \varepsilon_{(n)} + 2\varepsilon_{(n)}^2 + (y_{(n)}^\delta)^2 - y_{(n)}^2 \geqslant 0 \qquad \forall n \in \Lambda_N. \tag{5.11}$$

Now since, using $\|y\|_{L^2} = \left\|(y_{(n)})\right\|_{\ell^2}$, we get as in (5.5) that

$$\left|y_{(n)}^2 - (y_{(n)}^\delta)^2\right| \leqslant \delta\left(2\,\|y\|_{L^2} + \delta\right),$$

and hence, it follows that for inequality (5.11) to be satisfied, it suffices to have

$$(x^\dagger_{(n)})^2 - 4\left|x^\dagger_{(n)}\right|\left|\varepsilon_{(n)}\right| - \delta\left(2\left\|y\right\|_{L^2} + \delta\right) \geqslant 0, \qquad \forall\, n \in \Lambda_N.$$

However, since $\left|\varepsilon_{(n)}\right| \leqslant \left\|\varepsilon\right\|_{L^2} = \left\|x - x^\dagger\right\| \leqslant \left\|x - x_0\right\| + \left\|x_0 - x^\dagger\right\| \leqslant 7\rho$, this immediately follows from (5.9), which completes the proof. □

Remark. Similarly to the previous example, condition (5.3) is satisfied given that

$$\min_{n \in \Lambda_N}\left\{(x^\dagger_{(n)})^2\right\} \geqslant 28\left\|x^\dagger\right\|_{L^2}\rho + \bar{\delta}\left(\left\|y\right\|_{L^2} + \bar{\delta}\right),$$

which can always be satisfied given that $\left|x^\dagger_{(n)}\right| > 0$ for all $n \in \{1, \ldots, M\}$.

Remark. Note that one could also consider F as an operator from $H^1(0,1) \to L^2(0,1)$, in which case the local convexity of Φ^δ is still satisfied. Since, as noted in section 4, weak convergence in $H^1(0,1)$ implies strong convergence in $L^2(0,1)$, the convergence analysis carried out in the previous section then implies strong subsequential $L^2(0,1)$ convergence of the iterates x_k^δ of (1.22) to an element $\tilde{x} \in \mathcal{S}$ from the solution set.

Example 5.2. For this example, we consider the auto-convolution problem with exact solution $x^\dagger(s) := 10 + \sqrt{2}\sin(2\pi s)$. It follows that

$$x^\dagger_{(n)} = \left\langle x^\dagger, e_{(n)}\right\rangle = \begin{cases} 10, & n = 0, \\ 1, & n = 1, \\ 0, & \text{else} \end{cases}$$

and therefore, the convexity condition (5.9) simplifies to the following two inequalities

$$100 \geqslant 280\rho + \bar{\delta}\left(2\left\|y\right\|_{L^2} + \bar{\delta}\right), \qquad 1 \geqslant 28\rho + \bar{\delta}\left(2\left\|y\right\|_{L^2} + \bar{\delta}\right).$$

Hence, for the noise-free case (i.e. $\bar{\delta} = 0$) the functional Φ^0 is convex in $\mathcal{B}_{6\rho}(x_0)$ given that $\rho \leqslant 1/28 \approx 0.036$ and that $x_0 \in \mathcal{B}_\rho(x^\dagger)$, which is for example the case for the choice $x_0 = 10 + \frac{27}{28}\sqrt{2}\sin(2\pi s)$.

For discretizing the problem, we choose a uniform discretization of the interval $[0,1]$ into $N = 32$ equally spaced subintervals and introduce the standard finite element hat functions $\{\psi_i\}_{i=0}^N$ on this subdivision, which we use to discretize both $\mathcal{X}$ and $\mathcal{Y}$. Following the idea used in [26], we discretize F by the finite dimensional operator

$$F_N(x)(s) := \sum_{i=0}^N f_i(x)\psi_i(s), \qquad \text{where} \qquad f_i(x) := \int_0^1 x\left(\frac{i}{N} - t\right)x(t)\,\mathrm{d}t. \tag{5.12}$$

For computing the coefficients $f_i(x)$, we employ a 4-point Gaussian quadrature rule on each of the subintervals to approximate the integral in (5.12).

Now we again compare method (1.22) with (1.4). This time, the estimated scaling parameter has the value $\omega = 0.005$ and random noise with a relative noise level of 0.01% was added to the data. Again the discrepancy principle (1.6) with $\tau = 1$ was used and the proximal operator $\mathrm{prox}_{\omega\Psi}(.)$ in (1.22) was dropped. The results of the experiments, computed in MATLAB,

Table 2. Comparison of Landweber iteration (1.4) and its Nesterov accelerated version (1.22) when applied to the auto-convolution problem considered in example 5.2 for the choice $x^\dagger(s) = 10 + \sqrt{2}\sin(2\pi s)$ and $x_0(s) = 10 + \frac{27}{28}\sqrt{2}\sin(2\pi s)$ (left table) and $x^\dagger(s) = 10 + \sqrt{2}\sin(8\pi s)$ and $x_0(s) = 10 + \sqrt{2}\sin(2\pi s)$ (right table).

Method	k_*	Time (s)	$\lVert x^\dagger - x_k^\delta \rVert / \lVert x^\dagger \rVert$ (%)
Landweber	526	57	0.0244
Nesterov	50	6	0.0271

Method	k_*	Time (s)	$\lVert x^\dagger - x_k^\delta \rVert / \lVert x^\dagger \rVert$ (%)
Landweber	10 000	1067	9.57
Nesterov	797	87	0.65
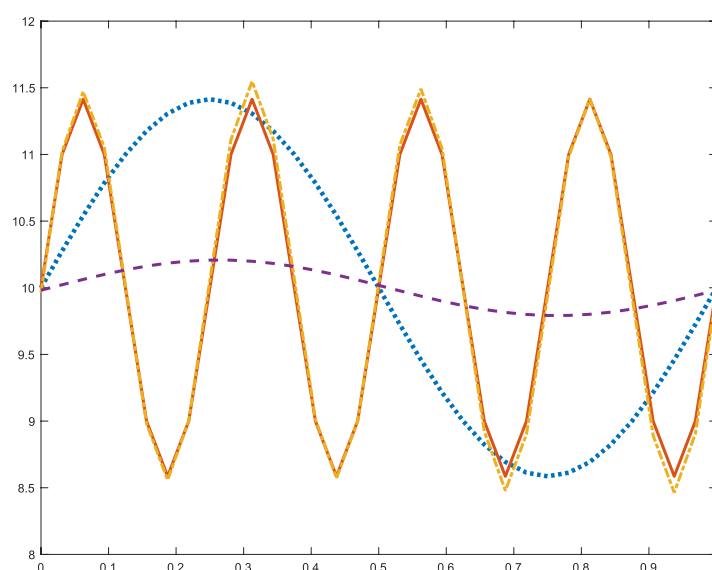
Figure 1. Auto-convolution example: initial guess x_0 (blue, dotted), exact solution $x^\dagger$ (red, solid), Landweber (1.4) reconstruction (purple, dashed), Nesterov (1.22) reconstruction (yellow, dash-dotted).

are displayed in the left part of table 2. Again the results clearly illustrate the advantages of Nesterov's acceleration strategy, which substantially decreases the required number of iterations and computational time, while leading to a relative error of essentially the same size as Landweber iteration.

The initial guess x_0 used for the experiment above is quite close to the exact solution $x^\dagger$. Although this is necessary for being able to guarantee convergence by our developed theory, it is not very practical. Hence, we want to see what happens if the solution and the initial guess are so far apart that they are no longer within the guaranteed area of convexity. For this, we consider the choice of $x^\dagger(s) = 10 + \sqrt{2}\sin(8\pi s)$ and $x_0(s) = 10 + \sqrt{2}\sin(2\pi s)$. The result can be seen in the right part of table 2. Landweber iteration was stopped after 10 000 iterations without having reached the discrepancy principle since no more progress was visible numerically. Consequently, it is clearly outperformed by (1.22), which manages to converge already

after 797 iterations, and with a much better relative error. The resulting reconstructions, depicted in figure 1, once again underline the usefulness of (1.22).

As an interesting remark, note that it seems that for the second example Landweber iteration gets stuck in a local minimum, while (1.22), after staying at this minimum for a while, manages to escape it, which is likely due to the combination step in (1.22).

5.3. Further examples

Besides the two rather academic examples presented above, we would like to cite a number of other examples where methods like (1.18) and (1.22) were successfully used, even though the key assumption of local convexity is not always known to hold for them.

First of all, in [18] the parameter estimation problem of magnetic resonance advection imaging (MRAI) was solved using a method very similar to (1.22). In MRAI, one aims at estimating the spatially varying pulse wave velocity (PWV) in blood vessels in the brain from magnetic resonance imaging (MRI) data. The PWV is directly connected to the health of the blood vessels and hence, it is used as a prognostic marker for various diseases in medical examinations. The data sets in MRAI are very large, making the direct application of second order methods like (1.9) or (1.10) difficult. However, since methods like (1.22) can deal with those large datasets, they were used in [18] for reconstructions of the PWV.

Secondly, in [17], numerical examples for various TPG methods (1.19), including the iteration (1.18), were presented. Among those is an example based on the imaging technique of single photon emission computed tomography (SPECT). Various numerical tests show that among all tested TPG methods, the method (1.18) clearly outperforms the rest, even though the local convexity assumption is not known to hold in this case. This is also demonstrated on an example based on a nonlinear Hammerstein operator.

Thirdly, method (1.18) was used in [19] to solve a problem in quantitative elastography, namely the reconstruction of the spatially varying Lamé parameters from full internal static displacement field measurements. Method (1.18) was used to obtain all reconstruction results presented in that paper, since ordinary first-order methods like Landweber iteration (1.4) were too slow to satisfy the demands required in practise.

Finally, in the numerical examples presented in [21], method (1.18) was used to accelerate the employed gradient/Kaczmarz methods. Furthermore, a convergence analysis of (1.18) for linear ill-posed problems including numerical examples is given in [28].

ORCID iDs

Simon Hubmer ⬤ https://orcid.org/0000-0002-8494-5188
Ronny Ramlau ⬤ https://orcid.org/0000-0002-0277-9615

References

[1] Anzengruber S W, Bürger S, Hofmann B and Steinmeyer G 2016 Variational regularization of complex deautoconvolution and phase retrieval in ultrashort laser pulse characterization *Inverse Problems* **32** 035002

[2] Attouch H and Peypouquet J 2016 The rate of convergence of Nesterov's accelerated forward–backward method is actually faster than $o(1/k^2)$ *SIAM J. Optim.* **26** 1824–34

[3] Bauschke H H and Combettes P L 2017 *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Berlin: Springer)

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

[5] Birkholz S, Steinmeyer G, Koke S, Gerth D, Bürger S and Hofmann B 2015 Phase retrieval via regularization in self-diffraction-based spectral interferometry *J. Opt. Soc. Am.* B **32** 983–92

[6] Blaschke B, Neubauer A and Scherzer O 1997 On convergence rates for the iteratively regularized Gauss–Newton method *IMA J. Numer. Anal.* **17** 421

[7] Bürger S and Hofmann B 2015 About a deficit in low-order convergence rates on the example of autoconvolution *Appl. Anal.* **94** 477–93

[8] Bürger S, Flemming J and Hofmann B 2016 On complex-valued deautoconvolution of compactly supported functions with sparse Fourier representation *Inverse Problems* **32** 104006

[9] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)

[10] Fleischer G and Hofmann B 1996 On inversion rates for the autoconvolution equation *Inverse Problems* **12** 419

[11] Gerth D, Hofmann B, Birkholz S, Koke S and Steinmeyer G 2014 Regularization of an autoconvolution problem in ultrashort laser pulse characterization *Inverse Problems Sci. Eng.* **22** 245–66

[12] Gorenflo R and Hofmann B 1994 On autoconvolution and regularization *Inverse Problems* **10** 353

[13] Hanke M 1991 Accelerated Landweber iterations for the solution of ill-posed equations *Numer. Math.* **60** 341–73

[14] Hanke M 1997 A regularizing Levenberg–Marquardt scheme, with applications to inverse groundwater filtration problems *Inverse Problems* **13** 79

[15] Hanke M, Neubauer A and Scherzer O 1995 A convergence analysis of the Landweber iteration for nonlinear ill-posed problems *Numer. Math.* **72** 21–37

[16] Hofmann B and Scherzer O 1998 Local ill-posedness and source conditions of operator equations in Hilbert spaces *Inverse Problems* **14** 1189

[17] Hubmer S and Ramlau R 2017 Convergence analysis of a two-point gradient method for nonlinear ill-posed problems *Inverse Problems* **33** 095004

[18] Hubmer S, Neubauer A, Ramlau R and Voss H U 2018 On the parameter estimation problem of magnetic resonance advection imaging *Inverse Problems Imaging* **12** 175–204

[19] Hubmer S, Sherina E, Neubauer A and Scherzer O 2018 Lamé parameter estimation from static displacement field measurements in the framework of nonlinear inverse problems *SIAM J. Imaging Sci.* **11** 1268–93

[20] Jin Q 2010 On a regularized Levenberg–Marquardt method for solving nonlinear inverse problems *Numer. Math.* **115** 229–59

[21] Jin Q 2016 Landweber–Kaczmarz method in Banach spaces with inexact inner solvers *Inverse Problems* **32** 104005

[22] Jin Q and Tautenhahn U 2009 On the discrepancy principle for some Newton type methods for solving nonlinear inverse problems *Numer. Math.* **111** 509–58

[23] Kaltenbacher B, Neubauer A and Scherzer O 2008 *Iterative Regularization Methods for Nonlinear Ill-Posed Problems* (Berlin: de Gruyter)

[24] Kindermann S 2017 Convergence of the gradient method for ill-posed problems *Inverse Problems Imaging* **11** 703–20

[25] Nesterov Y 1983 A method of solving a convex programming problem with convergence rate $O(1/k^2)$ *Sov. Math.—Dokl.* **27** 372–6

[26] Neubauer A 2000 On Landweber iteration for nonlinear ill-posed problems in Hilbert scales *Numer. Math.* **85** 309–28

[27] Neubauer A 2016 Some generalizations for Landweber iteration for nonlinear ill-posed problems in Hilbert scales *J. Inverse Ill-Posed Problems* **24** 393–406

[28] Neubauer A 2017 On Nesterov acceleration for Landweber iteration of linear ill-posed problems *J. Inverse Ill-Posed Problems* **25** 381–90

[29] Ramlau R 2003 TIGRA—an iterative algorithm for regularizing nonlinear ill-posed problems *Inverse Problems* **19** 433

[30] Ramlau R 1999 A modified Landweber method for inverse problems *Numer. Funct. Anal. Optim.* **20** 79–98

[31] Rockafellar R T, Wets M and Wets T J B 2009 *Variational Analysis* (*Grundlehren der Mathematischen Wissenschaften*) (Berlin: Springer)

[32] Scherzer O 1995 Convergence criteria of iterative methods based on Landweber iteration for solving nonlinear problems *J. Math. Anal. Appl.* **194** 911–33