Adaptive Nesterov momentum method for solving ill-posed inverse problems

OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **41** (2025) 025005 (34pp) https://doi.org/10.1088/1361-6420/ada8d3

Adaptive Nesterov momentum method for solving ill-posed inverse problems

Qinian Jin

Mathematical Sciences Institute, Australian National University, Canberra ACT 2601, Australia

E-mail: qinian.jin@anu.edu.au

Received 12 July 2024; revised 12 December 2024
Accepted for publication 10 January 2025
Published 23 January 2025

Abstract

Nesterov's acceleration strategy is renowned in speeding up the convergence of gradient-based optimization algorithms and has been crucial in developing fast first order methods for well-posed convex optimization problems. Although Nesterov's accelerated gradient method has been adapted as an iterative regularization method for solving ill-posed inverse problems, no general convergence theory is available except for some special instances. In this paper, we develop an adaptive Nesterov momentum method for solving ill-posed inverse problems in Banach spaces, where the step-sizes and momentum coefficients are chosen through adaptive procedures with explicit formulas. Additionally, uniform convex regularization functions are incorporated to detect the features of sought solutions. Under standard conditions, we establish the regularization property of our method when terminated by the discrepancy principle. Various numerical experiments demonstrate that our method outperforms the Landweber-type method in terms of the required number of iterations and the computational time.

Keywords: ill-posed inverse problems, adaptive Nesterov momentum method, the discrepancy principle, convergence

1. Introduction

Ill-posed inverse problems are prevalent in various scientific and engineering fields, where the goal is to determine an unknown quantity from indirect, incomplete, or noisy measurements.

An ill-posed problem is characterized by the lack of stability, meaning that small changes in the input data can lead to large variations in the solution. This instability makes direct solutions challenging and necessitates the use of specialized iterative algorithms to obtain meaningful results; see [6, 24, 28, 32, 34]. In this paper we will tackle ill-posed inverse problems by developing an adaptive Nesterov momentum method with fast convergence.

Consider ill-posed inverse problems governed by the operator equation of the form

$$F(x) = y, \tag{1}$$

where $F : \mathrm{dom}(F) \subset X \to Y$ is an operator between two Banach spaces X and Y with domain $\mathrm{dom}(F)$. We assume (1) has a solution, i.e. $y \in \mathrm{Ran}(F)$, the range of F. In practical scenarios, some *a priori* feature information on the sought solution is often available. It is important to integrate such information into the algorithm design. Let $\mathcal{R} : X \to (-\infty, \infty]$ be a proper, lower semi-continuous, p-convex function for some $p > 1$ that takes into account the available feature information. We pick $(x_0, \xi_0) \in X \times X^*$ with $\xi_0 \in \partial\mathcal{R}(x_0)$ as an initial guess, where X^* denotes the dual space of X and $\partial\mathcal{R}$ denotes the subdifferential of $\mathcal{R}$. Our aim is to determine a solution of (1) such that

$$D_{\mathcal{R}}^{\xi_0}\left(x^\dagger, x_0\right) = \min\left\{D_{\mathcal{R}}^{\xi_0}(x, x_0) : F(x) = y\right\}, \tag{2}$$

where $D_{\mathcal{R}}^{\xi_0}(x, x_0)$ denotes the Bregman distance induced by $\mathcal{R}$ at x_0 in the direction ξ_0; see section 2.

In practical applications, data are acquired through experiments and thus the exact data may not be available; instead we only have measurement data contaminated by noise. Let y^δ be a noisy data satisfying

$$\|y^\delta - y\| \leqslant \delta$$

with a known noise level $\delta > 0$. Based on the available noisy data, the Landweber-type method for solving ill-posed problems has been discussed in [4, 10, 15, 16, 20, 25, 33] and in particular the method of the form

$$\begin{aligned} \xi_{k+1}^\delta &= \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right), \\ x_{k+1}^\delta &= \arg\min_{x \in X}\left\{\mathcal{R}(x) - \langle \xi_{k+1}^\delta, x\rangle\right\} \end{aligned} \tag{3}$$

has been considered in [15, 20] for solving (1), where $\alpha_k^\delta > 0$ is the step-size, $J_s^Y : Y \to Y^*$ is the duality mapping of Y with the gauge function $t \to t^{s-1}$ for some $s > 1$, and $\{L(x) : x \in \mathrm{dom}(F)\}$ is a family of properly chosen bounded linear operators. In case F is Fréchect differentiable at x, we may take $L(x)$ to be the Fréchet derivative of F at x; otherwise, $L(x)$ should be appropriately chosen to substitute the non-existent Fréchet derivative. This method has been thoroughly analyzed in [15, 20] when it is terminated by the discrepancy principle

$$\|F\left(x_{k_\delta}^\delta\right) - y^\delta\| \leqslant \tau\delta < \|F\left(x_k^\delta\right) - y\|, \quad 0 \leqslant k < k_\delta \tag{4}$$

for some $\tau > 1$ and the regularization property has been established when the step-size is chosen by

$$\alpha_k^\delta = \min\left\{\frac{\mu_0\|F\left(x_k^\delta\right) - y^\delta\|^{(p-1)s}}{\|L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right)\|^p}, \mu_1\|F\left(x_k^\delta\right) - y^\delta\|^{p-s}\right\} \tag{5}$$

with suitable parameters μ_0 and μ_1. Due to its simplicity of implementation and low computational complexity per iteration, Landweber-type method is a popular choice for solving ill-posed inverse problems.

It is known that Landweber-type method is a slowly convergent method. As alternatives, second order iterative methods, such as the Levenberg–Maquardt method [9, 21], the iteratively regularized Gauss–Newton method [3, 22], and the nonstationary iterated Tikhonov regularization [7, 23], have been considered for solving ill-posed problems. Although these methods require less number of iterations to satisfy the respective stopping rules than the Landweber-type method, they always require more computational time in dealing with each iteration step. Consequently, the overall performance of these methods is even worse than the Landweber-type method for moderate and/or large size problems. Therefore, accelerating the Landweber-type method while preserving its straightforward implementation feature becomes significantly interesting.

Nesterov's acceleration strategy was proposed in [29] to speed up the convergence of gradient-based optimization algorithms. It has played a vital role on the development of fast first order methods for solving well-posed convex optimization problems [1, 2]. Based on Nesterov's acceleration strategy, an accelerated version of Landweber-type method has been proposed in [15] and a refined version of the method takes the form (see also [16, 36])

$$\begin{aligned} \xi_k^\delta &= \theta_k^\delta + \frac{k-1}{k+\gamma}\left(\theta_k^\delta - \theta_{k-1}^\delta\right), \\ z_k^\delta &= \arg\min_{x\in X}\left\{\mathcal{R}(x) - \langle \xi_k^\delta, x\rangle\right\}, \\ \theta_{k+1}^\delta &= \xi_k^\delta - \alpha_k^\delta L\left(z_k^\delta\right)^* J_s\left(F\left(z_k^\delta\right) - y^\delta\right) \end{aligned} \tag{6}$$

with $\gamma \geqslant 2$ and a suitable step sizes α_k^δ; after the iteration is terminated by a suitable stopping rule to output a stopping index k_δ, we then use

$$x_{k_\delta}^\delta = \arg\min_{x\in X}\left\{\mathcal{R}(x) - \langle \theta_{k_\delta}^\delta, x\rangle\right\}, \tag{7}$$

as an approximate solution. Actually, when Y is a Hilbert space, $s = 2$, and F is a bounded linear operator, the corresponding method of (6) and (7) can be derived by applying Nesterov's accelerated gradient method to the dual problem of (2), see [16]. The numerical results presented in [15] demonstrate the striking acceleration effect of the method (6). Inspired by the numerical observations in [15], further interest has been sparked on Nesterov acceleration method as seen in [14, 26, 27, 30]. The analysis of (6) however is extremely challenging and no efficient tools have been developed to analyze it in its general form. To the best of our knowledge, the following results are the only available ones on the method (6) in the context of regularizing ill-posed problems:

- When X and Y are Hilbert spaces, $F : X \to Y$ is a bounded linear operator and $\mathcal{R}(x) = \|x\|^2/2$, the corresponding method takes the form

$$z_k^\delta = x_k^\delta + \frac{k-1}{k+\gamma}\left(x_k^\delta - x_{k-1}^\delta\right), \quad x_{k+1}^\delta = z_k^\delta - \alpha F^*\left(Fz_k^\delta - y^\delta\right)$$

  with $x_{-1}^\delta = x_0^\delta = 0$ and $\gamma \geqslant 2$. When $0 < \alpha < 1/\|F\|^2$, the regularization property and the order optimality of the method have been established in [26, 30] under either *a priori* stopping rules or the discrepancy principle

$$\|Fx_{k_\delta}^\delta - y^\delta\| \leqslant \tau\delta < \|Fx_k^\delta - y^\delta\|, \quad 0 \leqslant k < k_\delta \tag{8}$$

with $\tau > 1$. Note that the implementation of this method under (8) requires to evaluate not only $Fz_k^\delta - y^\delta$ but also $Fx_k^\delta - y^\delta$ at each iteration step. Recently, it has been shown in [27] that the same convergence and convergence rate results as in [26] can be established when the method is terminated by the alternative discrepancy principle

$$\|Fz_{k_\delta}^\delta - y^\delta\| \leqslant \tau\delta < \|Fz_k^\delta - y^\delta\|, \quad 0 \leqslant k < k_\delta \tag{9}$$

which can save one-third computational time per iteration.

- When X is a Banach space, Y is a Hilbert space, and $F : X \to Y$ is a bounded linear operator, the method (6) has been considered in [16]. It has been shown in [16, theorem 3.8] that if the sought solution $x^\dagger$ satisfies the benchmark source condition

$$F^*\lambda^\dagger \in \partial\mathcal{R}\left(x^\dagger\right) \text{ for some } \lambda^\dagger \in Y,$$

then

$$\|x_{k_\delta}^\delta - x^\dagger\| = O\left(\delta^{1/2}\right)$$

if the stopping index k_δ is chosen such that $k_\delta \sim \delta^{-1/2}$. This demonstrates the acceleration effect over the Landweber iteration because the latter usually requires $O(\delta^{-1})$ iterations to achieve the same convergence rate. However, if no source condition is assumed, no convergence is yet available. Furthermore, nothing is known when the discrepancy principle is used to terminate the iteration.

Given the fast convergence of the method (6) observed in numerical experiments and the challenges in analyzing it theoretically, it is natural to consider modifying (6) to achieve provable convergence while maintaining its efficiency. By modifying the first equation in (6) by

$$\xi_k^\delta = \theta_k^\delta + \beta_k^\delta\left(\theta_k^\delta - \theta_{k-1}^\delta\right) \tag{10}$$

with a momentum coefficient β_k^δ to be determined, a two-point gradient method has been proposed in [13] for inverse problems in Hilbert spaces and then extended in [36] for inverse problems in Banach spaces. This method requires to determine at each iteration step a value $\beta_k^\delta > 0$ such that

$$\left(\left(\beta_k^\delta\right)^2 + \beta_k^\delta\right)\|\theta_k^\delta - \theta_{k-1}^\delta\|^2 - c\alpha_k^\delta\|F\left(z_k^\delta\right) - y^\delta\|^s \leqslant 0, \tag{11}$$

where $c > 0$ is a constant arising from the analysis. Note that z_k^δ depends on β_k^δ, so one can not simply solve (11) to determine β_k^δ. Instead, a backtracking line search procedure is required to find a value of β_k^δ that satisfies (11), see [13]. However, running a backtracking line search necessitates multiple evaluations of the forward operator at each iteration step, which incurs additional computational time and thus slows down the convergence.

In this paper we will consider a variant of (6) with the first equation replaced by (10) and with α_k^δ and β_k^δ being chosen by adaptive procedures; these α_k^δ and β_k^δ are given by explicit formulae and thus a backtracking line search procedure can be avoided. A notable difference is that, instead of using $x_{k_\delta}^\delta$ given by (7), we will use z_k^δ as approximate solutions. By renaming

z_k^δ as x_k^δ and rearranging the order of equations in (6), it leads to the following iteration scheme

$$\begin{aligned} \theta_{k+1}^\delta &= \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right), \\ \xi_{k+1}^\delta &= \theta_{k+1}^\delta + \beta_k^\delta\left(\theta_{k+1}^\delta - \theta_k^\delta\right), \\ x_{k+1}^\delta &= \arg\min_{x\in X}\left\{\mathcal{R}(x) - \langle \xi_{k+1}^\delta, x\rangle\right\}, \end{aligned} \tag{12}$$

where $(x_0^\delta, \xi_0^\delta) := (x_0, \xi_0) \in X \times X^*$ with $\xi_0 \in \partial\mathcal{R}(x_0)$ is an initial guess, and $\theta_0^\delta := \xi_0^\delta$. In order to determine α_k^δ and β_k^δ such that the method (12) has fast convergence property, we will consider the $\mathcal{R}$-induced Bregman distance between x_k^δ and a solution $\hat{x}$ of (1) and choose α_k^δ and β_k^δ such that this quantity to be as small as possible. Unfortunately this Bregman distance involves a solution $\hat{x}$ which is unknown, we can not directly minimizing this quantity to obtain α_k^δ and β_k^δ in general. Instead we will derive a suitable upper bound of this quantity and then minimize this upper bound to produce α_k^δ and β_k^δ successively. This leads us to propose the adaptive Nesterov momentum method for solving (1) using noisy data. We further demonstrate that our method is a well-defined regularization method. Extensive numerical simulations illustrate that our adaptive Nesterov momentum method accelerates the Landweber-type method by significantly reducing the number of iterations and the computational time.

Recently, an adaptive heavy ball method has been proposed in [17] to solve ill-posed inverse problems. This iterative method takes the form

$$\begin{aligned} \xi_{k+1}^\delta &= \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right) + \beta_k^\delta\left(\xi_k^\delta - \xi_{k-1}^\delta\right), \\ x_{k+1}^\delta &= \arg\min_{x\in X}\left\{\mathcal{R}(x) - \langle \xi_{k+1}^\delta, x\rangle\right\} \end{aligned} \tag{13}$$

with the step-size α_k^δ and the momentum coefficient β_k^δ being chosen adaptively. This method is a modification of (3) by adding the momentum term $\beta_k^\delta(\xi_k^\delta - \xi_{k-1}^\delta)$ to achieve acceleration. Our adaptive Nesterov momentum method is different from the method (13). Indeed, for the method (12) it is easy to see that

$$\begin{aligned} \xi_{k+1}^\delta &= \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right) + \beta_k^\delta\left(\theta_{k+1}^\delta - \theta_k^\delta\right) \\ &= \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right) + \beta_k^\delta\left(\xi_k^\delta - \xi_{k-1}^\delta\right) \\ &\quad - \beta_k^\delta\left(\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right) - \alpha_{k-1}^\delta L\left(x_{k-1}^\delta\right)^* J_s^Y\left(F\left(x_{k-1}^\delta\right) - y^\delta\right)\right). \end{aligned}$$

Comparing with the update of ξ_{k+1}^δ in (13), the update of ξ_{k+1}^δ in (12) relies on the extra term

$$-\beta_k^\delta\left(\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(F\left(x_k^\delta\right) - y^\delta\right) - \alpha_{k-1}^\delta L\left(x_{k-1}^\delta\right)^* J_s^Y\left(F\left(x_{k-1}^\delta\right) - y^\delta\right)\right)$$

which distinguishes the method (12) from (13).

This paper is organized as follows. In section 2 we will give some preliminaries from convex analysis and Banach spaces. In section 3 we will provide the detailed procedure for determining the step-size α_k^δ and the momentum coefficient β_k^δ, describe the adaptive Nesterov momentum method and show that the method is well-defined. In section 4 we will further demonstrate that our adaptive Nesterov momentum method is indeed a regularization method for solving ill-posed inverse problems. Finally in section 5 we will provide various numerical results to demonstrate the acceleration effect and the superior performance of our method over the Landweber type method.

2. Preliminaries

In this section we collect some necessary concepts and properties related to Banach spaces and convex analysis; for more details please refer to [5, 31, 35].

Throughout the paper we will use the same notation $\|\cdot\|$ to denote a norm in any Banach space; it should be clear from the context that which space is alluded to. Let X be a Banach space, we use X^* to denote its dual space. Given $x \in X$ and $\xi \in X^*$, we write $\langle \xi, x\rangle = \xi(x)$ for the duality pairing. For a bounded linear operator $A : X \to Y$ between two Banach spaces X and Y, we use $\mathrm{Ran}(A)$ and $A^* : Y^* \to X^*$ to denote its range and adjoint respectively. The modulus of smoothness of a Banach space X is the function

$$\rho_X(t) := \sup\{\|x+ty\| + \|x-ty\| - 2 : \|x\| = \|y\| = 1\}, \quad t \geqslant 0.$$

If $\lim_{t \searrow 0} \frac{\rho_X(t)}{t} = 0$, then X is called uniformly smooth. For any $1 < s < \infty$ the set-valued mapping $J_s^X : X \rightrightarrows X^*$ on a Banach space X defined by

$$J_s^X(x) := \left\{\xi \in X^* : \|\xi\| = \|x\|^{s-1} \text{ and } \langle \xi, x\rangle = \|x\|^s\right\} \tag{14}$$

is called the duality mapping of X with gauge function $t \to t^{s-1}$. By the Hahn–Banach extension theorem, $J_s^X(x) \neq \emptyset$ for each $x \in X$. If X is uniformly smooth, then the duality mapping J_s^X for any $1 < s < \infty$ is single valued and is norm-to-norm continuous from X to X^*.

Given a convex function $f : X \to (-\infty, \infty]$, we use

$$\mathrm{dom}(f) := \{x \in X : f(x) < \infty\}$$

to denote its effective domain. If $\mathrm{dom}(f) \neq \emptyset$, then f is called proper. The subdifferential of f is the set-valued mapping $\partial f : X \rightrightarrows X^*$ defined by

$$\partial f(x) := \{\xi \in X^* : f(\bar{x}) - f(x) - \langle \xi, \bar{x} - x\rangle \geqslant 0 \text{ for all } \bar{x} \in X\}$$

for each $x \in X$. The domain of ∂f is defined by

$$\mathrm{dom}(\partial f) := \{x \in \mathrm{dom}(f) : \partial f(x) \neq \emptyset\}.$$

For $x \in \mathrm{dom}(\partial f)$ each element $\xi \in \partial f(x)$ is called a subgradient of f at x. According to Asplund theorem, for $1 < s < \infty$ the subdifferential of the convex function $f_s(x) := \|x\|^s/s$ on a Banach space X is exactly the duality mapping J_s^X, i.e.

$$\partial f_s(x) = J_s^X(x), \quad \forall x \in X.$$

The Bregman distance induced by a proper convex function $f : X \to (-\infty, \infty]$ at $x \in \mathrm{dom}(\partial f)$ in a direction $\xi \in \partial f(x)$ is defined by

$$D_f^{\xi}(\bar{x}, x) := f(\bar{x}) - f(x) - \langle \xi, \bar{x} - x\rangle, \quad \forall \bar{x} \in X$$

which is always nonnegative and satisfies the identity

$$D_f^{\xi_2}(x, x_2) - D_f^{\xi_1}(x, x_1) = D_f^{\xi_2}(x_1, x_2) + \langle \xi_2 - \xi_1, x_1 - x\rangle \tag{15}$$

for all $x \in \mathrm{dom}(f), x_1, x_2 \in \mathrm{dom}(\partial f)$, and $\xi_1 \in \partial f(x_1)$, $\xi_2 \in \partial f(x_2)$.

A proper convex function $f: X \to (-\infty, \infty]$ is called p-convex for some $p > 1$ if there exists a constant $\sigma > 0$ such that

$$f(t\bar{x} + (1-t)x) + \sigma t(1-t)\|x - \bar{x}\|^p \leqslant tf(\bar{x}) + (1-t)f(x) \tag{16}$$

for all $\bar{x}, x \in \mathrm{dom}(f)$ and $t \in [0,1]$. It is straightforward to show that if f is p-convex then

$$D_f^{\xi}(\bar{x}, x) \geqslant \sigma\|x - \bar{x}\|^p \tag{17}$$

for all $\bar{x} \in X$, $x \in \mathrm{dom}(\partial f)$ and $\xi \in \partial f(x)$.

For a proper, lower semi-continuous, convex function $f: X \to (-\infty, \infty]$, its convex conjugate defined by

$$f^*(\xi) := \sup_{x \in X} \{\langle \xi, x\rangle - f(x)\}, \quad \xi \in X^*$$

is also a proper, lower semi-continuous, convex function and

$$\xi \in \partial f(x) \Longleftrightarrow x \in \partial f^*(\xi) \Longleftrightarrow f(x) + f^*(\xi) = \langle \xi, x\rangle. \tag{18}$$

Moreover, if f is p-convex with $p > 1$ then it follows from [35, corollary 3.5.11] that $\mathrm{dom}(f^*) = X^*$, f^* is Fréchet differentiable, its gradient ∇f^* maps X^* to X and

$$\|\nabla f^*(\xi_1) - \nabla f^*(\xi_2)\| \leqslant \left(\frac{\|\xi_1 - \xi_2\|}{2\sigma}\right)^{\frac{1}{p-1}}, \quad \forall \xi_1, \xi_2 \in X^*. \tag{19}$$

Consequently, it follows from (18) that

$$x = \nabla f^*(\xi) \Longleftrightarrow \xi \in \partial f(x) \Longleftrightarrow x = \arg\min_{z \in X} \{f(z) - \langle \xi, z\rangle\}. \tag{20}$$

Furthermore, for any $x, \bar{x} \in \mathrm{dom}(\partial f)$, $\xi \in \partial f(x)$ and $\bar{\xi} \in \partial f(\bar{x})$ it follows from (18) and (19) that

$$D_f^{\xi}(\bar{x}, x) = f^*(\xi) - f^*(\bar{\xi}) - \langle \xi - \bar{\xi}, \nabla f^*(\bar{\xi})\rangle \leqslant \frac{1}{p^*(2\sigma)^{p^*-1}}\|\xi - \bar{\xi}\|^{p^*}, \tag{21}$$

where $p^* := p/(p-1)$ is the number conjugate to p.

3. The adaptive Nesterov momentum method

In this section we will consider the method (12) and provide a motivation on how to choose α_k^{δ} and β_k^{δ} adaptively. We then show that the resulting algorithm is well-defined. We will work under the following standard conditions.

Assumption 1. $\mathcal{R}: X \to (-\infty, \infty]$ is a proper, lower semi-continuous, p-convex function for some $p > 1$ in the sense that there is a constant $\sigma > 0$ such that

$$\mathcal{R}(t\bar{x} + (1-t)x) + \sigma t(1-t)\|\bar{x} - x\|^p \leqslant t\mathcal{R}(\bar{x}) + (1-t)\mathcal{R}(x)$$

for all $\bar{x}, x \in \mathrm{dom}(\mathcal{R})$ and $0 \leqslant t \leqslant 1$.

Assumption 2. (i) X is a general Banach space, Y is a uniform smooth Banach spaces.

(ii) There exists $\rho>0$ such that $B_{2\rho}(x_0)\subset \mathrm{dom}(F)$ and (1) has a solution $\bar{x}$ such that $D_{\mathcal{R}}^{\xi_0}(\bar{x},x_0)\leqslant \sigma\rho^p$.

(iii) There is a family of bounded linear operators $\{L(x): X\to Y\}_{x\in B_{2\rho}(x_0)}$ such that $x\to L(x)$ is continuous on $B_{2\rho}(x_0)$ and there is $0\leqslant\eta<1$ such that

$$\|F(x)-F(\bar{x})-L(\bar{x})(x-\bar{x})\|\leqslant\eta\|F(x)-F(\bar{x})\|$$

for all $x,\bar{x}\in B_{2\rho}(x_0)$. Moreover, there is a constant $L>0$ such that $\|L(x)\|\leqslant L$ for all $x\in B_{2\rho}(x_0)$.

Under assumptions 1 and 2, it is easy to show that the problem (2) has a unique solution [20, lemma 3.2] which is denoted by $x^\dagger$.

In order for the method (12) to have nice approximation property, we need to choose α_k^δ and β_k^δ carefully. We set $\beta_{-1}^\delta=0$ and $\theta_{-1}^\delta=\theta_0^\delta=\xi_0^\delta$, then $\xi_0^\delta=\theta_0^\delta+\beta_{-1}^\delta(\theta_0^\delta-\theta_{-1}^\delta)$. Note that, by the last equation in (12) and the equation (20), we have $\xi_k^\delta\in\partial\mathcal{R}(x_k^\delta)$ for all integers $k\geqslant 0$. Let $\hat{x}$ denote any solution of (1) in $B_{2\rho}(x_0)\cap\mathrm{dom}(\mathcal{R})$, we consider the Bregman distance

$$\Delta_k^\delta:=D_{\mathcal{R}}^{\xi_k^\delta}\left(\hat{x},x_k^\delta\right)=\mathcal{R}(\hat{x})-\mathcal{R}\left(x_k^\delta\right)-\langle\xi_k^\delta,\hat{x}-x_k^\delta\rangle$$

to measure the approximation accuracy of x_k^δ to $\hat{x}$. It follows from (15), assumption 1 and (21) that

$$\begin{aligned}\Delta_{k+1}^\delta-\Delta_k^\delta&=D_{\mathcal{R}}^{\xi_{k+1}^\delta}\left(x_k^\delta,x_{k+1}^\delta\right)+\langle\xi_{k+1}^\delta-\xi_k^\delta,x_k^\delta-\hat{x}\rangle\\&\leqslant\frac{1}{p^*(2\sigma)^{p^*-1}}\|\xi_{k+1}^\delta-\xi_k^\delta\|^{p^*}+\langle\xi_{k+1}^\delta-\xi_k^\delta,x_k^\delta-\hat{x}\rangle.\end{aligned}$$

To carry out the analysis in a succinct manner, we set

$$r_k^\delta:=F\left(x_k^\delta\right)-y^\delta.$$

From the first two equations in (12) it follows that

$$\xi_{k+1}^\delta=\xi_k^\delta-\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)+\beta_k^\delta\left(\theta_{k+1}^\delta-\theta_k^\delta\right).$$

Therefore

$$\begin{aligned}\Delta_{k+1}^\delta-\Delta_k^\delta\leqslant&\frac{1}{p^*(2\sigma)^{p^*-1}}\left\|\beta_k^\delta\left(\theta_{k+1}^\delta-\theta_k^\delta\right)-\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\right\|^{p^*}\\&+\left\langle\beta_k^\delta\left(\theta_{k+1}^\delta-\theta_k^\delta\right)-\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right),x_k^\delta-\hat{x}\right\rangle.\end{aligned}\tag{22}$$

In order to produce efficient α_k^δ and β_k^δ, we need to estimate the right hand side of (22) as tight as possible. To this end, we first consider the special case that X is a Hilbert space and $p=2$, we then consider the general case.

When X is a Hilbert space and $p=2$, we can obtain from (22) that

$$\begin{aligned}\Delta_{k+1}^\delta-\Delta_k^\delta\leqslant&\frac{1}{4\sigma}\left\|\beta_k^\delta\left(\theta_{k+1}^\delta-\theta_k^\delta\right)-\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\right\|^2\\&+\left\langle\beta_k^\delta\left(\theta_{k+1}^\delta-\theta_k^\delta\right)-\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right),x_k^\delta-\hat{x}\right\rangle.\end{aligned}$$

By using the polarization identity in Hilbert spaces to treat the first term on the right hand side, we then have

$$\begin{aligned}\Delta_{k+1}^{\delta}-\Delta_{k}^{\delta} \leqslant & \frac{1}{4\sigma}\left(\alpha_{k}^{\delta}\right)^{2}\left\|L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right)\right\|^{2}+\frac{1}{4\sigma}\left(\beta_{k}^{\delta}\right)^{2}\left\|\theta_{k+1}^{\delta}-\theta_{k}^{\delta}\right\|^{2} \\ & -\frac{1}{2\sigma}\alpha_{k}^{\delta}\beta_{k}^{\delta}\left\langle L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right), \theta_{k+1}^{\delta}-\theta_{k}^{\delta}\right\rangle \\ & -\alpha_{k}^{\delta}\left\langle J_{s}^{Y}\left(r_{k}^{\delta}\right), L\left(x_{k}^{\delta}\right)\left(x_{k}^{\delta}-\hat{x}\right)\right\rangle \\ & +\beta_{k}^{\delta}\left\langle \theta_{k+1}^{\delta}-\theta_{k}^{\delta}, x_{k}^{\delta}-\hat{x}\right\rangle .\end{aligned}$$

Assume $x_{k}^{\delta} \in B_{2\rho}(x_{0})$. By using $\|y^{\delta}-y\| \leqslant \delta$, the property of duality mapping and assumption 2 (iii), we can obtain

$$\begin{aligned}-\left\langle J_{s}^{Y}\left(r_{k}^{\delta}\right), L\left(x_{k}^{\delta}\right)\left(x_{k}^{\delta}-\hat{x}\right)\right\rangle &= -\left\langle J_{s}^{Y}\left(r_{k}^{\delta}\right), F\left(x_{k}^{\delta}\right)-y^{\delta}\right\rangle-\left\langle J_{s}\left(r_{k}^{\delta}\right), y^{\delta}-y\right\rangle \\ &\quad -\left\langle J_{s}^{Y}\left(r_{k}^{\delta}\right), y-F\left(x_{k}^{\delta}\right)-L\left(x_{k}^{\delta}\right)\left(\hat{x}-x_{k}^{\delta}\right)\right\rangle \\ &\leqslant -\left\|r_{k}^{\delta}\right\|^{s}+\delta\left\|r_{k}^{\delta}\right\|^{s-1}+\eta\left\|r_{k}^{\delta}\right\|^{s-1}\left\|y-F\left(x_{k}^{\delta}\right)\right\| \\ &\leqslant -(1-\eta)\left\|r_{k}^{\delta}\right\|^{s}+(1+\eta)\delta\left\|r_{k}^{\delta}\right\|^{s-1}. \end{aligned} \tag{23}$$

Consequently

$$\begin{aligned}\Delta_{k+1}^{\delta}-\Delta_{k}^{\delta} \leqslant & \frac{1}{4\sigma}\left(\alpha_{k}^{\delta}\right)^{2}\left\|L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right)\right\|^{2}-(1-\eta)\alpha_{k}^{\delta}\left\|r_{k}^{\delta}\right\|^{s}+(1+\eta)\alpha_{k}^{\delta}\delta\left\|r_{k}^{\delta}\right\|^{s-1} \\ & +\frac{1}{4\sigma}\left(\beta_{k}^{\delta}\right)^{2}\left\|\theta_{k+1}^{\delta}-\theta_{k}^{\delta}\right\|^{2}-\frac{1}{2\sigma}\alpha_{k}^{\delta}\beta_{k}^{\delta}\left\langle L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right), \theta_{k+1}^{\delta}-\theta_{k}^{\delta}\right\rangle \\ & +\beta_{k}^{\delta}\left\langle \theta_{k+1}^{\delta}-\theta_{k}^{\delta}, x_{k}^{\delta}-\hat{x}\right\rangle .\end{aligned} \tag{24}$$

In order to have a fast convergent method, it is natural to find α_{k}^{δ} and β_{k}^{δ} such that the right hand side of (24) to be as small as possible. We can not minimize the right hand side of (24) with respect to α_{k}^{δ} and β_{k}^{δ} simultaneously because the definition of θ_{k+1}^{δ} requires the knowledge of α_{k}^{δ}. Therefore, we will first determine how to choose α_{k}^{δ} and then use α_{k}^{δ} to make a choice of β_{k}^{δ}.

Note that when $\beta_{k}^{\delta}=0$, all the terms involving β_{k}^{δ} on the right hand side of (24) drop and only the first three terms survive, i.e. the right hand side of (24) becomes

$$\frac{1}{4\sigma}\left(\alpha_{k}^{\delta}\right)^{2}\left\|L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right)\right\|^{2}-\left((1-\eta)\left\|r_{k}^{\delta}\right\|-(1+\eta)\delta\right)\alpha_{k}^{\delta}\left\|r_{k}^{\delta}\right\|^{s-1}. \tag{25}$$

In order to guarantee this quantity to be negative, we need to choose α_{k}^{δ} such that

$$0 \leqslant \alpha_{k}^{\delta}<\frac{4\sigma\left((1-\eta)\left\|r_{k}^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_{k}^{\delta}\right\|^{s-1}}{\left\|L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right)\right\|^{2}}$$

in case

$$\left\|r_{k}^{\delta}\right\|>\frac{1+\eta}{1-\eta}\delta. \tag{26}$$

This motivates us to choose α_{k}^{δ} as

$$\alpha_{k}^{\delta}:=\min\left\{\frac{\mu_{0}\left((1-\eta)\left\|r_{k}^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_{k}^{\delta}\right\|^{s-1}}{\left\|L\left(x_{k}^{\delta}\right)^{*} J_{s}^{Y}\left(r_{k}^{\delta}\right)\right\|^{2}}, \mu_{1}\left\|r_{k}^{\delta}\right\|^{2-s}\right\} \tag{27}$$

when (26) holds, where $0 < \mu_0 < 4\sigma$ and $\mu_1 > 0$ are two preassigned numbers; the appearance of $\mu_1 \|r_k^\delta\|^{2-s}$ is used to enhance the stability. We remark that one may consider choosing α_k^δ to minimize (25) over the interval $[0, \mu_1 \|r_k^\delta\|^{2-s}]$ which consequently leads to the choice of α_k^δ given by (27) with $\mu_0 = 2\sigma$. However, we will use (27) which has the flexibility of tuning μ_0 to achieve overall better performance during the whole iteration process.

By plugging the choice of α_k^δ from (27) into (24) we have

$$\begin{aligned}\Delta_{k+1}^\delta - \Delta_k^\delta \leqslant & -\left(1 - \frac{\mu_0}{4\sigma}\right) \alpha_k^\delta \left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right) \|r_k^\delta\|^{s-1} \\ & + \frac{1}{4\sigma} \left(\beta_k^\delta\right)^2 \|\theta_{k+1}^\delta - \theta_k^\delta\|^2 - \frac{1}{2\sigma} \alpha_k^\delta \beta_k^\delta \left\langle L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right), \theta_{k+1}^\delta - \theta_k^\delta \right\rangle \\ & + \beta_k^\delta \left\langle \theta_{k+1}^\delta - \theta_k^\delta, x_k^\delta - \hat{x} \right\rangle. \end{aligned} \tag{28}$$

Since α_k^δ has been chosen, we can calculate θ_{k+1}^δ. Next we determine how to choose β_k^δ. A natural idea is to choose β_k^δ such that the right hand side of (28) is minimized over some interval $[0, \hat{\beta}_k]$ with $0 < \hat{\beta}_k \leqslant \infty$ to be set by users. This gives

$$\beta_k^\delta = \min\left\{ \max\left\{ 0, \frac{\alpha_k^\delta \left\langle L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right), \theta_{k+1}^\delta - \theta_k^\delta \right\rangle - 2\sigma\gamma_{k+1}^\delta}{\|\theta_{k+1}^\delta - \theta_k^\delta\|^2} \right\}, \hat{\beta}_k \right\}$$

whenever $\theta_{k+1}^\delta \neq \theta_k^\delta$, where

$$\gamma_{k+1}^\delta := \left\langle \theta_{k+1}^\delta - \theta_k^\delta, x_k^\delta - \hat{x} \right\rangle. \tag{29}$$

However, this formula for β_k^δ is not computable because it involves γ_{k+1}^δ which requires knowledge of the unknown solution $\hat{x}$. We need to replace the term γ_{k+1}^δ with a suitable surrogate. According to the first two equations in (12) we have

$$\theta_{k+1}^\delta = \xi_k^\delta - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right) = \theta_k^\delta + \beta_{k-1}^\delta \left(\theta_k^\delta - \theta_{k-1}^\delta\right) - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right).$$

Therefore

$$\begin{aligned}\gamma_{k+1}^\delta &= \left\langle \beta_{k-1}^\delta \left(\theta_k^\delta - \theta_{k-1}^\delta\right) - \alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right), x_k^\delta - \hat{x} \right\rangle \\ &= \beta_{k-1}^\delta \left\langle \theta_k^\delta - \theta_{k-1}^\delta, x_k^\delta - x_{k-1}^\delta \right\rangle + \beta_{k-1}^\delta \left\langle \theta_k^\delta - \theta_{k-1}^\delta, x_{k-1}^\delta - \hat{x} \right\rangle \\ &\quad - \alpha_k^\delta \left\langle J_s^Y\left(r_k^\delta\right), L\left(x_k^\delta\right)\left(x_k^\delta - \hat{x}\right) \right\rangle. \end{aligned}$$

By virtue of (23) we then have

$$\begin{aligned}\gamma_{k+1}^\delta \leqslant & \beta_{k-1}^\delta \left\langle \theta_k^\delta - \theta_{k-1}^\delta, x_k^\delta - x_{k-1}^\delta \right\rangle + \beta_{k-1}^\delta \gamma_k^\delta - (1-\eta)\alpha_k^\delta \|r_k^\delta\|^s \\ & + (1+\eta)\delta\alpha_k^\delta \|r_k^\delta\|^{s-1}. \end{aligned} \tag{30}$$

This motivates us to define $\{\tilde{\gamma}_k^\delta\}$ by setting $\tilde{\gamma}_0^\delta = 0$ and

$$\begin{aligned}\tilde{\gamma}_{k+1}^\delta = & \beta_{k-1}^\delta \left\langle \theta_k^\delta - \theta_{k-1}^\delta, x_k^\delta - x_{k-1}^\delta \right\rangle + \beta_{k-1}^\delta \tilde{\gamma}_k^\delta - (1-\eta)\alpha_k^\delta \|r_k^\delta\|^s \\ & + (1+\eta)\delta\alpha_k^\delta \|r_k^\delta\|^{s-1} \end{aligned} \tag{31}$$

for $k \geqslant 0$ once x_l^δ for $0 \leqslant l \leqslant k$, α_l^δ for $0 \leqslant l \leqslant k$, and β_l^δ for $0 \leqslant l \leqslant k-1$ are defined. Therefore, by using $\tilde{\gamma}_{k+1}^\delta$ to replace γ_{k+1}^δ, it leads us to propose the following adaptive Nesterov

momentum method for solving ill-posed inverse problem (1) when X is a Hilbert space and $p=2$.

Algorithm 1 (adaptive Nesterov momentum method: X is Hilbertian and $p=2$). Take $\xi_0 \in X$, calculate $x_0 := \arg\min_{x\in X}\{\mathcal{R}(x) - \langle \xi_0, x\rangle\}$, and set $x_{-1}^\delta = x_0^\delta = x_0$, $\theta_0^\delta = \xi_0^\delta = \xi_0$. Pick $\tau > 1$, $\mu_0 > 0$, $\mu_1 > 0$, $s > 1$, and a user defined sequence $\{\hat\beta_k\}$ of positive numbers. Let $\beta_{-1}^\delta = 0$, $\tilde\gamma_0^\delta = 0$ and $m_0^\delta = 0$. For $k \geqslant 0$ do the following:

(i) Calculate $r_k^\delta := F(x_k^\delta) - y^\delta$. If $\|r_k^\delta\| \leqslant \tau\delta$, stop and output x_k^δ;
(ii) Calculate $g_k^\delta := L(x_k^\delta)^* J_s^Y(r_k^\delta)$, choose α_k^δ as in (27), i.e.

$$\alpha_k^\delta := \min\left\{ \frac{\mu_0\left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)\|r_k^\delta\|^{s-1}}{\|g_k^\delta\|^2}, \mu_1\|r_k^\delta\|^{2-s}\right\};$$

(iii) Determine $\tilde\gamma_{k+1}^\delta$ by

$$\begin{aligned}\tilde\gamma_{k+1}^\delta = &-(1-\eta)\alpha_k^\delta\|r_k^\delta\|^s + (1+\eta)\delta\alpha_k^\delta\|r_k^\delta\|^{s-1}\\ &+ \beta_{k-1}^\delta\langle m_k^\delta, x_k^\delta - x_{k-1}^\delta\rangle + \beta_{k-1}^\delta\tilde\gamma_k^\delta;\end{aligned}$$

(iv) Update θ_{k+1}^δ by $\theta_{k+1}^\delta = \xi_k^\delta - \alpha_k^\delta g_k^\delta$, calculate $m_{k+1}^\delta := \theta_{k+1}^\delta - \theta_k^\delta$ and determine β_k^δ by

$$\beta_k^\delta = \begin{cases} \min\left\{\max\left\{0, \frac{\alpha_k^\delta\langle g_k^\delta, m_{k+1}^\delta\rangle - 2\sigma\tilde\gamma_{k+1}^\delta}{\|m_{k+1}^\delta\|^2}\right\}, \hat\beta_k\right\} & \text{if } m_{k+1}^\delta \neq 0,\\ 0 & \text{if } m_{k+1}^\delta = 0;\end{cases}$$

(v) Update ξ_{k+1}^δ and x_{k+1}^δ by

$$\xi_{k+1}^\delta = \theta_{k+1}^\delta + \beta_k^\delta m_{k+1}^\delta, \quad x_{k+1}^\delta = \arg\min_{x\in X}\left\{\mathcal{R}(x) - \langle \xi_{k+1}^\delta, x\rangle\right\}.$$

Before presenting the analysis, let us give some remarks on algorithm 1 concerning the produced momentum coefficients and its computational complexity compared with the Landweber type method (3).

Remark 3.1. The momentum coefficient β_k^δ produced by algorithm 1 is always nonnegative. We would like to point out that algorithm 1 actually can produce $\beta_k^\delta > 0$ frequently. To see this, it suffices to show that if $\beta_{k-1}^\delta = 0$ for some $k \geqslant 0$ with $\|r_k^\delta\| > \tau\delta$, then it must hold $\beta_k^\delta > 0$. Indeed, when $\beta_{k-1}^\delta = 0$, from the description of algorithm 1 we can see that

$$\tilde\gamma_{k+1}^\delta = -\left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)\alpha_k^\delta\|r_k^\delta\|^{s-1} < 0$$

and

$$m_{k+1}^\delta = \theta_{k+1}^\delta - \theta_k^\delta = \xi_k^\delta - \theta_k^\delta - \alpha_k^\delta g_k^\delta = \beta_{k-1}^\delta m_k^\delta - \alpha_k^\delta g_k^\delta = -\alpha_k^\delta g_k^\delta;$$

consequently, by using the choice of α_k^δ, we have

$$\begin{aligned}\alpha_k^\delta\langle g_k^\delta, m_{k+1}^\delta\rangle - 2\sigma\tilde\gamma_{k+1}^\delta &= -\left(\alpha_k^\delta\right)^2\|g_k^\delta\|^2 + 2\sigma\left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)\alpha_k^\delta\|r_k^\delta\|^{s-1}\\ &\geqslant (2\sigma - \mu_0)\left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)\alpha_k^\delta\|r_k^\delta\|^{s-1} > 0\end{aligned}$$

if $0<\mu_0<2\sigma$. Moreover, by using the property of duality mapping, assumption 2, and $\|r_k^\delta\|>\tau\delta$ with $\tau>(1+\eta)/(1-\eta)$, we have

$$\begin{aligned}\langle g_k^\delta, x_k^\delta-\hat{x}\rangle &= \langle J_s^Y\left(r_k^\delta\right), L\left(x_k^\delta\right)\left(x_k^\delta-\hat{x}\right)\rangle\\ &= \langle J_s^Y\left(r_k^\delta\right), r_k^\delta+y^\delta-y+y-F\left(x_k^\delta\right)-L\left(x_k^\delta\right)\left(\hat{x}-x_k^\delta\right)\rangle\\ &\geqslant \|r_k^\delta\|^s-\delta\|r_k^\delta\|^{s-1}-\eta\|r_k^\delta\|^{s-1}\|F\left(x_k^\delta\right)-y\|\\ &\geqslant \left((1-\eta)\|r_k^\delta\|-(1+\eta)\delta\right)\|r_k^\delta\|^{s-1}>0\end{aligned}$$

which implies $g_k^\delta\neq 0$ and hence $m_{k+1}^\delta\neq 0$. Therefore, based on the above facts and the definition of β_k^δ it is easy to see that $\beta_k^\delta>0$. The frequent occurrence of $\beta_k^\delta>0$ contributes to a fast convergence of algorithm 1.

Remark 3.2. Note that, for the implementation of one step of algorithm 1, the most expensive parts are the calculations of r_k^δ, g_k^δ and x_{k+1}^δ which are common for the Landweber-type method (3); the computational load for other parts relating to α_k^δ, $\tilde{\gamma}_k^\delta$ and β_k^δ can be negligible. Therefore, the computational complexity per iteration of algorithm 1 is marginally higher than, but very close to, that of one step of the Landweber-type method.

In the design of algorithm 1, the discrepancy principle has been incorporated, i.e. the iteration stops as long as $\|r_k^\delta\|\leqslant\tau\delta$ is satisfied for the first time. In the following we will prove that algorithm 1 is well-defined by showing that the iteration indeed can stop in finite many steps.

Lemma 3.3. *Consider algorithm 1 and let $n\geqslant 0$ be an integer such that $x_k^\delta\in B_{2\rho}(x_0)$ for all $0\leqslant k<n$. Define γ_k^δ by (29) for $0\leqslant k\leqslant n$. Then $\gamma_k^\delta\leqslant\tilde{\gamma}_k^\delta$ for all $0\leqslant k\leqslant n$.*

Proof. Since $\gamma_0^\delta=\tilde{\gamma}_0^\delta=0$, the result is true for $k=0$. Assume next that $\gamma_k^\delta\leqslant\tilde{\gamma}_k^\delta$ for some $0\leqslant k<n$. By noting that $\beta_{k-1}^\delta\geqslant 0$, we may use (30) and the definition of $\tilde{\gamma}_{k+1}^\delta$ to obtain

$$\begin{aligned}\gamma_{k+1}^\delta &\leqslant \beta_{k-1}^\delta\left\langle m_k^\delta, x_k^\delta-x_{k-1}^\delta\right\rangle+\beta_{k-1}^\delta\gamma_k^\delta-(1-\eta)\alpha_k^\delta\|r_k^\delta\|^s+(1+\eta)\delta\alpha_k^\delta\|r_k^\delta\|^{s-1}\\ &\leqslant \beta_{k-1}^\delta\left\langle m_k^\delta, x_k^\delta-x_{k-1}^\delta\right\rangle+\beta_{k-1}^\delta\tilde{\gamma}_k^\delta-(1-\eta)\alpha_k^\delta\|r_k^\delta\|^s+(1+\eta)\delta\alpha_k^\delta\|r_k^\delta\|^{s-1}\\ &= \tilde{\gamma}_{k+1}^\delta.\end{aligned}$$

By the induction principle, this shows the result. □

Lemma 3.4. *Let assumptions 1 and 2 hold with X being a Hilbert space and $p=2$. Consider algorithm 1 with $0<\mu_0<4\sigma$ and $\tau>(1+\eta)/(1-\eta)$. Let $n\geqslant 0$ be an integer such that $\|r_k^\delta\|>\tau\delta$ for all $0\leqslant k<n$. Then $x_k^\delta\in B_{2\rho}(x_0)$ for all $0\leqslant k\leqslant n$ and*

$$\Delta_{k+1}^\delta\leqslant\Delta_k^\delta-c_0\alpha_k^\delta\|r_k^\delta\|^s$$

for all $0\leqslant k<n$, where $c_0:=(1-\frac{\mu_0}{4\sigma})(1-\eta-\frac{1+\eta}{\tau})$, $\Delta_k^\delta:=D_{\mathcal{R}}^{\xi_k^\delta}(\hat{x},x_k^\delta)$ and $\hat{x}$ denotes any solution of (1) in $B_{2\rho}(x_0)\cap\mathrm{dom}(\mathcal{R})$.

Proof. We first show by induction that

$$x_k^\delta\in B_{2\rho}\left(x_0\right)\quad\text{and}\quad D_{\mathcal{R}}^{\xi_k^\delta}\left(x^\dagger,x_k^\delta\right)\leqslant D_{\mathcal{R}}^{\xi_0}\left(x^\dagger,x_0\right)\tag{32}$$

for all integers $0\leqslant k\leqslant n$. It is trivial for $k=0$ as $\xi_0^\delta=\xi_0$ and $x_0^\delta=x_0$. Next we assume that (32) holds for all $0\leqslant k\leqslant l$ for some $l<n$. According to (28), $\beta_l\geqslant 0$, and lemma 3.3 we have for

any solution $\hat{x}$ of (1) in $B_{2\rho}(x_0)\cap \mathrm{dom}(\mathcal{R})$ that

$$\begin{aligned}\Delta_{l+1}^{\delta}-\Delta_l^{\delta} &\leqslant -\left(1-\frac{\mu_0}{4\sigma}\right)\alpha_l^{\delta}\left((1-\eta)\left\|r_l^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_l^{\delta}\right\|^{s-1}\\ &\quad+\frac{1}{4\sigma}\left(\beta_l^{\delta}\right)^2\left\|m_{l+1}^{\delta}\right\|^2-\frac{1}{2\sigma}\alpha_l^{\delta}\beta_l^{\delta}\left\langle g_l^{\delta},m_{l+1}^{\delta}\right\rangle+\beta_l^{\delta}\gamma_{l+1}^{\delta}\\ &\leqslant -\left(1-\frac{\mu_0}{4\sigma}\right)\alpha_l^{\delta}\left((1-\eta)\left\|r_l^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_l^{\delta}\right\|^{s-1}\\ &\quad+\frac{1}{4\sigma}\left(\beta_l^{\delta}\right)^2\left\|m_{l+1}^{\delta}\right\|^2-\frac{1}{2\sigma}\alpha_l^{\delta}\beta_l^{\delta}\left\langle g_l^{\delta},m_{l+1}^{\delta}\right\rangle+\beta_l^{\delta}\tilde{\gamma}_{l+1}^{\delta}.\end{aligned}$$

Note that β_l^{δ} is the minimizer of the function $t\to h_l(t)$ over $[0,\hat{\beta}_l]$, where

$$h_l(t):=\frac{1}{4\sigma}t^2\|m_{l+1}^{\delta}\|^2-\frac{1}{2\sigma}\alpha_l^{\delta}t\left\langle g_l^{\delta},m_{l+1}^{\delta}\right\rangle+t\tilde{\gamma}_{l+1}^{\delta},$$

we can conclude that

$$\begin{aligned}\Delta_{l+1}^{\delta}-\Delta_l^{\delta} &\leqslant -\left(1-\frac{\mu_0}{4\sigma}\right)\alpha_l^{\delta}\left((1-\eta)\left\|r_l^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_l^{\delta}\right\|^{s-1}+h_l\left(\beta_l^{\delta}\right)\\ &\leqslant -\left(1-\frac{\mu_0}{4\sigma}\right)\alpha_l^{\delta}\left((1-\eta)\left\|r_l^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_l^{\delta}\right\|^{s-1}+h_l(0)\\ &= -\left(1-\frac{\mu_0}{4\sigma}\right)\alpha_l^{\delta}\left((1-\eta)\left\|r_l^{\delta}\right\|-(1+\eta)\delta\right)\left\|r_l^{\delta}\right\|^{s-1}. \qquad (33)\end{aligned}$$

Since $\delta\leqslant\|r_l^{\delta}\|/\tau$, we can further obtain

$$\Delta_{l+1}^{\delta}-\Delta_l^{\delta}\leqslant -\left(1-\frac{\mu_0}{4\sigma}\right)\left(1-\eta-\frac{1+\eta}{\tau}\right)\alpha_l^{\delta}\|r_l^{\delta}\|^s=-c_0\alpha_l^{\delta}\|r_l^{\delta}\|^s. \qquad (34)$$

By virtue of this inequality with $\hat{x}=x^{\dagger}$, the induction hypothesis, we have

$$D_{\mathcal{R}}^{\xi_{l+1}^{\delta}}\left(x^{\dagger},x_{l+1}^{\delta}\right)\leqslant D_{\mathcal{R}}^{\xi_l^{\delta}}\left(x^{\dagger},x_l^{\delta}\right)\leqslant D_{\mathcal{R}}^{\xi_0}\left(x^{\dagger},x_0\right)\leqslant\sigma\rho^2$$

which together with the strong convexity of $\mathcal{R}$ implies that $\sigma\|x_{l+1}^{\delta}-x^{\dagger}\|^2\leqslant\sigma\rho^2$ and hence $\|x_{l+1}^{\delta}-x^{\dagger}\|\leqslant\rho$. Since $\|x^{\dagger}-x_0\|\leqslant\rho$, we thus have $\|x_{l+1}^{\delta}-x_0\|\leqslant 2\rho$, i.e. $x_{l+1}^{\delta}\in B_{2\rho}(x_0)$. We therefore complete the proof of (32). As a direct consequence, we can see that (34) holds for all $0\leqslant l<n$ which shows the desired result. □

Based on lemma 3.4 we now can prove that algorithm 1 is well-defined, i.e. iteration must terminate in finite many steps and the iterates always stay in $B_{2\rho}(x_0)$ until the iteration is terminated.

Theorem 3.5. *Let assumptions 1 and 2 hold with X being a Hilbert space and $p=2$. Consider algorithm 1 with noisy data y^{δ} satisfying $\|y^{\delta}-y\|\leqslant\delta$ with noise level $\delta>0$. Assume that $0<\mu_0<4\sigma$ and $\tau>(1+\eta)/(1-\eta)$. Then the algorithm must terminate in finite many steps, i.e. there exists a finite integer k_{δ} such that*

$$\|r_{k_{\delta}}^{\delta}\|\leqslant\tau\delta<\|r_k^{\delta}\|,\quad 0\leqslant k<k_{\delta}. \qquad (35)$$

Proof. Let $l \geqslant 0$ be an integer such that $\|r_k^\delta\| > \tau\delta$ for all $0 \leqslant k \leqslant l$. It then follows from lemma 3.4 that

$$c_0 \sum_{k=0}^{l} \alpha_k^\delta \|r_k^\delta\|^s \leqslant \sum_{k=0}^{l} \left(\Delta_k^\delta - \Delta_{k+1}^\delta\right) = \Delta_0^\delta - \Delta_{l+1}^\delta \leqslant \Delta_0^\delta = D_{\mathcal{R}}^{\xi_0}(\hat{x}, x_0).$$

According to the choice of α_k^δ, the property of duality mapping, and $\|L(x)\| \leqslant L$ for all $x \in B_{2\rho}(x_0)$, we can see

$$\alpha_k^\delta \geqslant \min\left\{\frac{\mu_0\left(1-\eta-(1+\eta)/\tau\right)\|r_k^\delta\|^s}{\|L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\|^2}, \mu_1\|r_k^\delta\|^{2-s}\right\} \geqslant c_1\|r_k^\delta\|^{2-s}$$

for all $0 \leqslant k \leqslant l$, where

$$c_1 := \min\left\{\frac{\mu_0}{L^2}\left(1-\eta-\frac{1+\eta}{\tau}\right), \mu_1\right\} > 0.$$

Thus

$$(l+1)\, c_0 c_1 \tau^2 \delta^2 \leqslant c_0 c_1 \sum_{k=0}^{l} \|r_k^\delta\|^2 \leqslant c_0 \sum_{k=0}^{l} \alpha_k^\delta \|r_k^\delta\|^s \leqslant D_{\mathcal{R}}^{\xi_0}(\hat{x}, x_0) < \infty.$$

If there is no finite integer k_δ such that (35) holds, then we can take l to be any positive integer. Letting $l \to \infty$ in the above equation gives a contradiction. Thus, the algorithm must terminate in finite many steps. □

Next we turn to consider the case that X is a general Banach space. We do not have the polarization identity to use any more. To deal with the first term on the right hand side of (22), we use the following elementary inequality

$$(a+b)^q \leqslant t\left(t^{\frac{1}{q-1}}-1\right)^{1-q} a^q + tb^q$$

for any $a, b \geqslant 0$, $q > 1$ and $t > 1$. Let us fix a number $t > 1$. From (22) and (23) it then follows that

$$\begin{aligned}
&\Delta_{k+1}^\delta - \Delta_k^\delta \\
&\leqslant \frac{1}{p^*(2\sigma)^{p^*-1}}\left(t\left(t^{\frac{1}{p^*-1}}-1\right)^{1-p^*}\|\beta_k^\delta\left(\theta_{k+1}^\delta - \theta_k^\delta\right)\|^{p^*} + t\|\alpha_k^\delta L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\|^{p^*}\right) \\
&\quad - \alpha_k^\delta\left\langle J_s^Y\left(r_k^\delta\right), L\left(x_k^\delta\right)\left(x_k^\delta - \hat{x}\right)\right\rangle + \beta_k^\delta\left\langle\theta_{k+1}^\delta - \theta_k^\delta, x_k^\delta - \hat{x}\right\rangle \\
&\leqslant \frac{t}{p^*(2\sigma)^{p^*-1}}\left(\alpha_k^\delta\right)^{p^*}\|L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\|^{p^*} - \left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)\alpha_k^\delta\|r_k^\delta\|^{s-1} \\
&\quad + \frac{t\left(t^{p-1}-1\right)^{1-p^*}}{p^*(2\sigma)^{p^*-1}}\left(\beta_k^\delta\right)^{p^*}\|\theta_{k+1}^\delta - \theta_k^\delta\|^{p^*} + \beta_k^\delta\left\langle\theta_{k+1}^\delta - \theta_k^\delta, x_k^\delta - \hat{x}\right\rangle. \qquad (36)
\end{aligned}$$

When (26) holds, by dropping the terms involving β_k^δ we may use the same reasoning as above to motivate the choice of α_k^δ as follows

$$\alpha_k^\delta = \min\left\{\mu_0\left(\frac{p^*}{t}\right)^{p-1}\frac{\left((1-\eta)\|r_k^\delta\| - (1+\eta)\delta\right)^{p-1}\|r_k^\delta\|^{(p-1)(s-1)}}{\|L\left(x_k^\delta\right)^* J_s^Y\left(r_k^\delta\right)\|^p}, \mu_1\|r_k^\delta\|^{p-s}\right\}, \qquad (37)$$

where $0<\mu_0<2\sigma$ and $\mu_1>0$. Once α_k^δ is selected, we may use the above same procedure to select β_k^δ as

$$\beta_k^\delta = \min\left\{\left(\max\left\{0, -\frac{\left(2\sigma\left(t^{p-1}-1\right)\right)^{p^*-1}\tilde{\gamma}_{k+1}^\delta}{t\|\theta_{k+1}^\delta-\theta_k^\delta\|^{p^*}}\right\}\right)^{p-1}, \hat{\beta}_k\right\} \tag{38}$$

with a user-defined $\hat{\beta}_k>0$, where $\tilde{\gamma}_{k+1}^\delta$ is defined by (31). Putting all these together leads to the following adaptive Nesterov momentum method for the general situation.

Algorithm 2 (adaptive Nesterov momentum method: general case). Take $\xi_0\in X^*$, calculate $x_0:=\arg\min_{x\in X}\{\mathcal{R}(x)-\langle\xi_0,x\rangle\}$, and set $x_{-1}^\delta=x_0^\delta=x_0$, $\theta_0^\delta=\xi_0^\delta=\xi_0$. Pick $\tau>1$, $\mu_0>0$, $\mu_1>0$, $s>1$, $t>1$, and a user defined sequence $\{\hat{\beta}_k\}$ of positive numbers. Let $\beta_{-1}^\delta=0$, $\tilde{\gamma}_0^\delta=0$ and $m_0^\delta=0$. For $k\geqslant 0$ do the following:

(i) Calculate $r_k^\delta:=F(x_k^\delta)-y^\delta$. If $\|r_k^\delta\|\leqslant\tau\delta$, stop and output x_k^δ;
(ii) Calculate $g_k^\delta:=L(x_k^\delta)^*J_s^Y(r_k^\delta)$, choose α_k^δ as in (37), and define $\tilde{\gamma}_{k+1}^\delta$ by (31).
(iii) Update θ_{k+1}^δ by $\theta_{k+1}^\delta=\xi_k^\delta-\alpha_k^\delta g_k^\delta$, calculate $m_{k+1}^\delta:=\theta_{k+1}^\delta-\theta_k^\delta$ and determine β_k^δ by (38).
(iv) Update ξ_{k+1}^δ and x_{k+1}^δ by

$$\xi_{k+1}^\delta=\theta_{k+1}^\delta+\beta_k^\delta m_{k+1}^\delta,\quad x_{k+1}^\delta=\arg\min_{x\in X}\left\{\mathcal{R}(x)-\langle\xi_{k+1}^\delta,x\rangle\right\}.$$

For algorithm 2 we may use the same argument for proving theorem 3.5 to obtain easily the following result.

Theorem 3.6. *Let assumptions 1 and 2 hold. Consider algorithm 2 with noisy data y^δ satisfying $\|y^\delta-y\|\leqslant\delta$ with noise level $\delta>0$. Assume that $0<\mu_0<2\sigma$ and $\tau>(1+\eta)/(1-\eta)$. Then the algorithm must output a finite integer k_δ satisfying $\|r_{k_\delta}^\delta\|\leqslant\tau\delta$ for the first time. Moreover $x_k^\delta\in B_{2\rho}(x_0)$ for all $0\leqslant k\leqslant k_\delta$ and there is a constant $c_0'>0$ such that*

$$\Delta_{k+1}^\delta\leqslant\Delta_k^\delta-c_0'\alpha_k^\delta\|r_k^\delta\|^s$$

for all $0\leqslant k<k_\delta$, where $\Delta_k^\delta:=D_{\mathcal{R}}^{\xi_k^\delta}(\hat{x},x_k^\delta)$ and $\hat{x}$ denotes any solution of (1) in $B_{2\rho}(x_0)\cap$ $\mathrm{dom}(\mathcal{R})$.

4. Convergence analysis

In this section we will show that our adaptive Nesterov momentum method is a regularization method for solving ill-posed problems. We will focus on algorithm 1; the same argument can be adapted easily to analyze algorithm 2. Let k_δ be the integer determined by the discrepancy principle (35), we will show that $x_{k_\delta}^\delta$ converges to a solution of (1) as $\delta\to 0$. To this end, we will consider the counterpart of algorithm 1 with the exact data and investigate its convergence behavior. We will then connect algorithm 1 with its counterpart for exact data by establishing a stability result. Based on these results, we will show the regularization property of algorithm 1 finally.

4.1. The method with exact data

We first consider the couterpart of algorithm 1 with exact data. For convenience we give the detailed formulation of the corresponding algorithm in the following.

Algorithm 3 (adaptive Nesterov momentum method with exact data: *X* is Hilbertian and *p* = 2). Take $\xi_0 \in X$, calculate $x_0 := \arg\min_{x\in X}\{\mathcal{R}(x) - \langle \xi_0, x\rangle\}$, and set $x_{-1} = x_0$, $\theta_0 = \xi_0$. Pick $\mu_0 > 0$, $\mu_1 > 0$, $s > 1$ and a user defined sequence $\{\hat{\beta}_k\}$ of positive numbers. Let $\beta_{-1} = 0$, $\tilde{\gamma}_0 = 0$ and $m_0 = 0$. For $k \geqslant 0$ do the following:

(i) Calculate $r_k := F(x_k) - y$ and $g_k := L(x_k)^* J_s^Y(r_k)$, choose α_k by

$$\alpha_k := \begin{cases} \min\left\{\frac{\mu_0(1-\eta)\|r_k\|^s}{\|g_k\|^2}, \mu_1\|r_k\|^{2-s}\right\}, & \text{if } r_k \neq 0, \\ 0, & \text{if } r_k = 0; \end{cases}$$

(ii) Determine $\tilde{\gamma}_{k+1}$ by

$$\tilde{\gamma}_{k+1} = -(1-\eta)\alpha_k\|r_k\|^s + \beta_{k-1}\langle m_k, x_k - x_{k-1}\rangle + \beta_{k-1}\tilde{\gamma}_k;$$

(iii) Update θ_{k+1} by $\theta_{k+1} = \xi_k - \alpha_k g_k$, calculate $m_{k+1} := \theta_{k+1} - \theta_k$ and determine β_k by

$$\beta_k = \begin{cases} \min\left\{\max\left\{0, \frac{\alpha_k\langle g_k, m_{k+1}\rangle - 2\sigma\tilde{\gamma}_{k+1}}{\|m_{k+1}\|^2}\right\}, \hat{\beta}_k\right\} & \text{if } m_{k+1} \neq 0, \\ 0 & \text{if } m_{k+1} = 0; \end{cases}$$

(iv) Update ξ_{k+1} and x_{k+1} by

$$\xi_{k+1} = \theta_{k+1} + \beta_k m_{k+1}, \quad x_{k+1} = \arg\min_{x\in X}\{\mathcal{R}(x) - \langle \xi_{k+1}, x\rangle\}.$$

By using the similar argument in the proof of lemma 3.4, for algorithm 3 we can easily obtain the following result which in particular demonstrates that $x_k \in B_{2\rho}(x_0)$ for all integers $k \geqslant 0$ and hence algorithm 3 is well-defined.

Lemma 4.1. *Let assumptions 1 and 2 hold with X being a Hilbert space and $p = 2$. Consider algorithm 3 with $0 < \mu_0 < 4\sigma$. Then $x_k \in B_{2\rho}(x_0)$ for all integers $k \geqslant 0$ and for any solution $\hat{x}$ of (1) in $B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$ there holds*

$$\Delta_{k+1} \leqslant \Delta_k - c_2\alpha_k\|r_k\|^s, \quad \forall k \geqslant 0,$$

where $\Delta_k := D_{\mathcal{R}}^{\xi_k}(\hat{x}, x_k)$ and $c_2 := \left(1 - \frac{\mu_0}{4\sigma}\right)(1-\eta) > 0$. Consequently $\{\Delta_k\}$ is monotonically decreasing and

$$\sum_{k=0}^{\infty}\alpha_k\|r_k\|^s < \infty.$$

In order to show the convergence of the sequence $\{x_k\}$ generated by algorithm 3, we need the following general convergence result; see [20, proposition 3.6] and [18, proposition 2.3].

Proposition 4.2. *Let assumptions 1 and 2 hold and consider the equation (1). Let $\{x_k\} \subset B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$ and $\{\xi_k\} \subset X^*$ be such that*

(i) *$\xi_k \in \partial\mathcal{R}(x_k)$ for all k;*

(ii) *for any solution* $\hat{x}$ *of* (1) *in* $B_{2\rho}(x_0)\cap \mathrm{dom}(\mathcal{R})$ *the sequence* $\{D_{\mathcal{R}}^{\xi_k}(\hat{x},x_k)\}$ *is convergent;*
(iii) $\lim_{k\to\infty}\|F(x_k)-y\|=0$.
(iv) *there is a subsequence* $\{k_n\}$ *of integers with* $k_n\to\infty$ *such that for any solution* $\hat{x}$ *of* (1) *in* $B_{2\rho}(x_0)\cap\mathrm{dom}(\mathcal{R})$ *there holds*

$$\lim_{l\to\infty}\sup_{n\geqslant l}|\langle \xi_{k_n}-\xi_{k_l}, x_{k_n}-\hat{x}\rangle|=0. \tag{39}$$

Then there exists a solution x^* *of* (1) *in* $B_{2\rho}(x_0)\cap\mathrm{dom}(\mathcal{R})$ *such that*

$$\lim_{k\to\infty} D_{\mathcal{R}}^{\xi_k}(x^*,x_k)=0.$$

If, in addition, $\xi_{k+1}-\xi_k\in\overline{\mathrm{Ran}(L(x^\dagger)^*)}$ *for all* k, *then* $x^*=x^\dagger$, *where* $x^\dagger$ *denotes the unique solution of* (1) *satisfying* (2).

Based on lemma 4.1 and proposition 4.2, we are now ready to show the convergence of algorithm 3.

Theorem 4.3. *Let assumptions* 1 *and* 2 *hold with* X *being a Hilbert space and* $p=2$. *Consider algorithm* 3 *with* $0<\mu_0<4\sigma$ *and* $\beta:=\sup_{k\geqslant 0}\hat{\beta}_k<1$. *Then there exists a solution* x^* *of* (1) *in* $B_{2\rho}(x_0)$ *such that*

$$\lim_{k\to\infty} D_{\mathcal{R}}^{\xi_k}(x^*,x_k)=0 \quad \text{and} \quad \lim_{k\to\infty}\|x_k-x^*\|=0. \tag{40}$$

If, in addition, $\mathrm{Ran}(L(x)^*)\subset\overline{\mathrm{Ran}(L(x^\dagger)^*)}$ *for all* $x\in B_{2\rho}(x_0)$, *then* $x^*=x^\dagger$.

Proof. We will use proposition 4.2 to prove the conclusion. The item (i) in proposition 4.2 holds automatically by the definition of x_k. According to lemma 4.1 we immediately obtain the item (ii) in proposition 4.2. To show item (iii), from the definition of $\{\alpha_k\}$ we note that, when $r_k\neq 0$, there holds

$$\alpha_k=\min\left\{\frac{\mu_0(1-\eta)\|r_k\|^s}{\|L(x_k)^* J_s^Y(r_k)\|^2},\mu_1\|r_k\|^{2-s}\right\}\geqslant c_3\|r_k\|^{2-s}$$

with $c_3:=\min\{\mu_0(1-\eta)/L^2,\mu_1\}$. Thus $\alpha_k\|r_k\|^s\geqslant c_3\|r_k\|^2$ for all $k\geqslant 0$. From lemma 4.1 it then follows that

$$c_3\sum_{k=0}^{\infty}\|r_k\|^2\leqslant\sum_{k=0}^{\infty}\alpha_k\|r_k\|^s<\infty. \tag{41}$$

This in particular implies (iii) in proposition 4.2, i.e.

$$\lim_{k\to\infty}\|r_k\|=0. \tag{42}$$

We now verify (iv) in proposition 4.2. By taking a sequence $\{\varepsilon_k\}$ of positive number such that $\sum_{k=1}^{\infty}\varepsilon_k<\infty$, we define

$$\Phi_k:=\|r_k\|^2+\varepsilon_k.$$

It then follows from (41) that

$$\sum_{k=0}^{\infty} \Phi_k < \infty. \tag{43}$$

Moreover, since $\{\Phi_k\}$ is a sequence of positive numbers satisfying $\Phi_k \to 0$ as $k \to \infty$, we can choose a strictly increasing sequence $\{k_n\}$ of integers such that $k_0 = 0$ and k_n, for each $n \geqslant 1$, is the first integer satisfying

$$k_n \geqslant k_{n-1} + 1 \quad \text{and} \quad \Phi_{k_n} \leqslant \Phi_{k_{n-1}}.$$

For this chosen $\{k_n\}$ it is easy to see that

$$\Phi_k \geqslant \Phi_{k_n}, \quad \forall 0 \leqslant k \leqslant k_n. \tag{44}$$

For the above chosen $\{k_n\}$, we now show (39) for any solution $\hat{x}$ of (1) in $B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$. For any integers $l < n$ we write

$$\langle \xi_{k_n} - \xi_{k_l}, x_{k_n} - \hat{x} \rangle = \sum_{k=k_l}^{k_n - 1} \langle \xi_{k+1} - \xi_k, x_{k_n} - \hat{x} \rangle. \tag{45}$$

From the description of algorithm 3, we have

$$\theta_{k+1} = \xi_k - \alpha_k g_k \quad \text{and} \quad \xi_{k+1} = \theta_{k+1} + \beta_k m_{k+1}, \quad \forall k \geqslant 0.$$

This implies that

$$\xi_{k+1} - \xi_k = -\alpha_k g_k + \beta_k m_{k+1} \quad \text{and} \quad m_{k+1} = -\alpha_k g_k + \beta_{k-1} m_k \tag{46}$$

for all integers $k \geqslant 0$. From the second equation in (46) and $m_0 = 0$ it follows that

$$m_{k+1} = -\alpha_k g_k - \sum_{i=0}^{k-1} \left(\prod_{l=i}^{k-1} \beta_l \right) \alpha_i g_i = -\sum_{i=0}^{k} \left(\prod_{l=i}^{k-1} \beta_l \right) \alpha_i g_i.$$

Combining this with the first equation in (46) gives

$$\xi_{k+1} - \xi_k = -\alpha_k g_k - \sum_{i=0}^{k} \left(\prod_{l=i}^{k} \beta_l \right) \alpha_i g_i. \tag{47}$$

Thus

$$\begin{aligned} \langle \xi_{k+1} - \xi_k, x_{k_n} - \hat{x} \rangle &= -\alpha_k \langle g_k, x_{k_n} - \hat{x} \rangle - \sum_{i=0}^{k} \left(\prod_{l=i}^{k} \beta_l \right) \alpha_i \langle g_i, x_{k_n} - \hat{x} \rangle \\ &= -\alpha_k \left\langle J_s^Y(r_k), L(x_k)(x_{k_n} - \hat{x}) \right\rangle \\ &\quad - \sum_{i=0}^{k} \left(\prod_{l=i}^{k} \beta_l \right) \alpha_i \left\langle J_s^Y(r_i), L(x_i)(x_{k_n} - \hat{x}) \right\rangle. \end{aligned}$$

Therefore, by using the Cauchy–Schwarz inequality and the property of duality mapping, we can obtain

$$\begin{aligned}&|\langle \xi_{k+1}-\xi_k, x_{k_n}-\hat{x}\rangle|\\ &\leqslant \alpha_k\|r_k\|^{s-1}\|L(x_k)(x_{k_n}-\hat{x})\|+\sum_{i=0}^{k}\left(\prod_{l=i}^{k}\beta_l\right)\alpha_i\|r_i\|^{s-1}\|L(x_i)(x_{k_n}-\hat{x})\|\end{aligned}$$

With the help of assumption 2 (iii) it is easy to obtain

$$\begin{aligned}\|L(x_i)(x_{k_n}-\hat{x})\| &\leqslant \|L(x_i)(x_i-\hat{x})\|+\|L(x_i)(x_{k_n}-x_i)\|\\ &\leqslant (1+\eta)\left(\|r_i\|+\|F(x_i)-F(x_{k_n})\|\right)\\ &\leqslant (1+\eta)\left(2\|r_i\|+\|r_{k_n}\|\right).\end{aligned}$$

Consequently

$$\begin{aligned}|\langle \xi_{k+1}-\xi_k, x_{k_n}-\hat{x}\rangle| &\leqslant (1+\eta)\,\alpha_k\|r_k\|^{s-1}\left(2\|r_k\|+\|r_{k_n}\|\right)\\ &\quad+(1+\eta)\sum_{i=0}^{k}\left(\prod_{l=i}^{k}\beta_l\right)\alpha_i\|r_i\|^{s-1}\left(2\|r_i\|+\|r_{k_n}\|\right).\end{aligned}$$

Since $\alpha_i\leqslant\mu_1\|r_i\|^{2-s}$ for all $i\geqslant 0$ and $0\leqslant\beta_l\leqslant\hat{\beta}_l\leqslant\beta<1$ for all $l\geqslant 0$, we can further obtain

$$\begin{aligned}&|\langle \xi_{k+1}-\xi_k, x_{k_n}-\hat{x}\rangle|\\ &\leqslant (1+\eta)\,\mu_1\|r_k\|\left(2\|r_k\|+\|r_{k_n}\|\right)+(1+\eta)\,\mu_1\sum_{i=0}^{k}\beta^{k+1-i}\|r_i\|\left(2\|r_i\|+\|r_{k_n}\|\right)\\ &\leqslant (1+\eta)\,\mu_1\Phi_k^{1/2}\left(2\Phi_k^{1/2}+\Phi_{k_n}^{1/2}\right)+(1+\eta)\,\mu_1\sum_{i=0}^{k}\beta^{k+1-i}\Phi_i^{1/2}\left(2\Phi_i^{1/2}+\Phi_{k_n}^{1/2}\right).\end{aligned}$$

By virtue of (44), for $k<k_n$ we then have

$$|\langle \xi_{k+1}-\xi_k, x_{k_n}-\hat{x}\rangle|\leqslant 3\,(1+\eta)\,\mu_1\Phi_k+3\,(1+\eta)\,\mu_1\sum_{i=0}^{k}\beta^{k+1-i}\Phi_i.$$

This together with (45) then gives

$$|\langle \xi_{k_n}-\xi_{k_l}, x_{k_n}-\hat{x}\rangle|\leqslant 3\,(1+\eta)\,\mu_1\sum_{k=k_l}^{k_n-1}\Phi_k+3\,(1+\eta)\,\mu_1\sum_{k=k_l}^{k_n-1}\sum_{i=0}^{k}\beta^{k+1-i}\Phi_i. \tag{48}$$

According to (43), the first term on the right hand side of (48) can be easily handled. We deal with the second term by showing that

$$\lim_{l\to\infty}\sup_{n\geqslant l}\left\{\sum_{k=k_l}^{k_n-1}\sum_{i=0}^{k}\beta^{k-i}\Phi_i\right\}=0. \tag{49}$$

To see this, we use $\beta < 1$ to obtain

$$\begin{aligned}\sum_{k=k_l}^{k_n-1}\sum_{i=0}^{k}\beta^{k-i}\Phi_i &= \sum_{k=k_l}^{k_n-1}\sum_{i=0}^{k_l-1}\beta^{k-i}\Phi_i + \sum_{k=k_l}^{k_n-1}\sum_{i=k_l}^{k}\beta^{k-i}\Phi_i \\ &= \sum_{i=0}^{k_l-1}\left(\sum_{k=k_l}^{k_n-1}\beta^{k-i}\right)\Phi_i + \sum_{i=k_l}^{k_n-1}\left(\sum_{k=i}^{k_n-1}\beta^{k-i}\right)\Phi_i \\ &\leqslant \frac{1}{1-\beta}\sum_{i=0}^{k_l-1}\beta^{k_l-i}\Phi_i + \frac{1}{1-\beta}\sum_{i=k_l}^{k_n-1}\Phi_i.\end{aligned}$$

Since $\Phi_i \to 0$ as $i \to \infty$, we can find a constant C such that $\Phi_i \leqslant C$ for all i. Thus

$$\begin{aligned}\sum_{i=0}^{k_l-1}\beta^{k_l-i}\Phi_i &= \sum_{i=0}^{[k_l/2]}\beta^{k_l-i}\Phi_i + \sum_{i=[k_l/2]+1}^{k_l-1}\beta^{k_l-i}\Phi_i \\ &\leqslant C\sum_{i=0}^{[k_l/2]}\beta^{k_l-i} + \sum_{i=[k_l/2]+1}^{k_l-1}\Phi_i \\ &\leqslant \frac{C}{1-\beta}\beta^{k_l-[k_l/2]} + \sum_{i=[k_l/2]+1}^{\infty}\Phi_i \\ &\to 0\end{aligned}$$

as $l \to \infty$ by the facts $0 \leqslant \beta < 1$ and (43). By (43) we also have $\sum_{i=k_l}^{k_n-1}\Phi_i \leqslant \sum_{i=k_l}^{\infty}\Phi_i \to 0$ as $l \to \infty$ and thus (49) follows. Consequently, it follows from (48) that

$$\sup_{n\geqslant l}|\langle \xi_{k_n} - \xi_{k_l}, x_{k_n} - \hat{x}\rangle| \to 0 \quad \text{as } l \to \infty$$

and hence (iv) in proposition 4.2 is verified. Therefore, we may use proposition 4.2 to conclude the existence of a solution x^* of (1) in $B_{2\rho}(x_0)\cap \mathrm{dom}(\mathcal{R})$ such that (40) holds.

If $\mathrm{Ran}(L(x)^*) \subset \overline{\mathrm{Ran}(L(x^\dagger)^*)}$ for all $x \in B_{2\rho}(x_0)$, then we may use (47) to obtain

$$\xi_{k+1} - \xi_k \in \mathrm{Ran}\left(L(x_k)^*\right) \oplus \cdots \oplus \mathrm{Ran}\left(L(x_0)^*\right) \subset \overline{\mathrm{Ran}\left(L(x^\dagger)^*\right)}.$$

Thus, we may use the last part of proposition 4.2 to conclude $x^* = x^\dagger$. □

4.2. Stability and regularization property

In this subsection we will show that algorithm 1 is indeed a regularization method for solving ill-posed problems. We need the following stability result which connects algorithm 1 with its counterpart for exact data, i.e. algorithm 3.

Lemma 4.4. *Let assumptions 1 and 2 hold with X being a Hilbert space and $p = 2$. Consider algorithm 1 with $0 < \mu_0 < 4\sigma$ and $\tau > (1+\eta)/(1-\eta)$ and let k_δ be the output integer. Consider also algorithm 3 with the same $\{\hat{\beta}_k\}$, μ_0, μ_1 and the initial guess ξ_0. Then there hold*

$$\theta_k^\delta \to \theta_k, \quad \xi_k^\delta \to \xi_k \quad \text{and} \quad x_k^\delta \to x_k \quad \text{as } \delta \to 0$$

for all integers $0 \leqslant k \leqslant N$, *where* $N := \liminf_{\delta\to 0} k_\delta$.

Proof. By setting $\alpha_{-1}^\delta = \alpha_{-1} = 0$, $g_{-1}^\delta = g_{-1} = 0$ and $r_{-1}^\delta = r_{-1} = 0$, we will use an induction argument to show that

$$\begin{aligned} &\theta_k^\delta \to \theta_k, \quad \alpha_{k-1}^\delta g_{k-1}^\delta \to \alpha_{k-1} g_{k-1}, \quad \alpha_{k-1}^\delta \|r_{k-1}^\delta\|^{s-1} \to \alpha_{k-1}\|r_{k-1}\|^{s-1}, \\ &\tilde\gamma_k^\delta \to \tilde\gamma_k, \quad \beta_{k-1}^\delta \tilde\gamma_k^\delta \to \beta_{k-1}\tilde\gamma_k, \quad \beta_{k-1}^\delta m_k^\delta \to \beta_{k-1} m_k, \quad \xi_k^\delta \to \xi_k, \quad x_k^\delta \to x_k \end{aligned} \tag{50}$$

as $\delta \to 0$ for all integers $0 \leqslant k \leqslant N$. The case $k = 0$ is trivial because $\theta_0^\delta = \theta_0 = \xi_0 = \xi_0^\delta, x_0^\delta = x_0$, $m_0^\delta = m_0 = 0$ and $\tilde\gamma_0^\delta = \tilde\gamma_0 = 0$. Now we assume that (50) holds for all $0 \leqslant k \leqslant l$ for some integer $0 \leqslant l < N$ and show that (50) holds as well for $k = l+1$. Note that $l + 1 \leqslant k_\delta$ for sufficiently small $\delta > 0$.

We first show that

$$\theta_{l+1}^\delta \to \theta_{l+1}, \quad \alpha_l^\delta g_l^\delta \to \alpha_l g_l, \quad \alpha_l^\delta \|r_l^\delta\|^{s-1} \to \alpha_l \|r_l\|^{s-1} \quad \text{as } \delta \to 0 \tag{51}$$

by considering two cases on r_l. If $r_l = 0$, then $g_l = 0$. By noting that $0 \leqslant \alpha_l^\delta \leqslant \mu_1 \|r_l^\delta\|^{2-s}$ and using the induction hypothesis, we have

$$0 \leqslant \alpha_l^\delta \|r_l^\delta\|^{s-1} \leqslant \mu_1 \|r_l^\delta\| \to \mu_1 \|r_l\| = 0$$

and

$$\|\alpha_l^\delta g_l^\delta\| \leqslant \mu_1 \|r_l^\delta\|^{2-s} \|L\left(x_l^\delta\right)^* J_s^Y\left(r_l^\delta\right)\| \leqslant \mu_1 L \|r_l^\delta\| \to \mu_1 L \|r_l\| = 0$$

as $\delta \to 0$. Thus the last two assertions in (51) hold. Furthermore, since $\theta_{l+1} = \xi_l$ and $\theta_{l+1}^\delta = \xi_l^\delta - \alpha_l^\delta g_l^\delta$, we may use the induction hypothesis and the result just established to derive that $\theta_{l+1}^\delta \to \xi_l - \alpha_l g_l = \xi_l = \theta_{l+1}$ as $\delta \to 0$. Thus (51) is confirmed for this case. If $r_l \neq 0$, then $g_l \neq 0$ because otherwise we would have

$$\begin{aligned} 0 = \langle g_l, x_l - \hat x\rangle &= \left\langle J_s^Y(r_l), L(x_l)(x_l - \hat x)\right\rangle \\ &= \left\langle J_s^Y(r_l), r_l + y - F(x_l) - L(x_l)(\hat x - x_l)\right\rangle \\ &\geqslant \|r_l\|^s - \|r_l\|^{s-1}\|y - F(x_l) - L(x_l)(\hat x - x_l)\| \\ &\geqslant (1-\eta)\|r_l\|^s > 0 \end{aligned}$$

which is a contradiction. Consequently, by the induction hypothesis and the continuity of F, J_s and $x \to L(x)$ we have $r_l^\delta \to r_l$ and $g_l^\delta \to g_l$ as $\delta \to 0$ and consequently $r_l^\delta \neq 0$ and $g_l^\delta \neq 0$ for small $\delta > 0$. Thus, by the definition of α_l^δ and α_l we can conclude $\alpha_l^\delta \to \alpha_l$ and hence the last two assertions in (51) hold. Moreover

$$\theta_{l+1}^\delta = \xi_l^\delta - \alpha_l^\delta g_l^\delta \to \xi_l - \alpha_l g_l = \theta_{l+1}$$

as $\delta \to 0$. We therefore establish (51) again.

Next we show that

$$\tilde\gamma_{l+1}^\delta \to \tilde\gamma_{l+1}, \quad \beta_l^\delta \tilde\gamma_{l+1}^\delta \to \beta_l \tilde\gamma_{l+1}, \quad \beta_l^\delta m_{l+1}^\delta \to \beta_l m_{l+1} \quad \text{as } \delta \to 0. \tag{52}$$

By the induction hypothesis and the definition of $\tilde\gamma_{l+1}^\delta$ and $\tilde\gamma_{l+1}$, we can immediately obtain $\tilde\gamma_{l+1}^\delta \to \tilde\gamma_{l+1}$ as $\delta \to 0$. To show the remaining two assertions in (52), we consider two cases on m_{l+1}. If $m_{l+1} \neq 0$, by using (51) we have $m_{l+1}^\delta \to m_{l+1}$ as $\delta \to 0$ and thus $m_{l+1}^\delta \neq 0$ for small

$\delta > 0$. Therefore, we may use the definition of β_l^δ and β_l and the second equation in (51) to conclude $\beta_l^\delta \to \beta_l$ as $\delta \to 0$ and thus the last two assertions in (52) follows. In the following we consider the case $m_{l+1} = 0$. This implies $\beta_l = 0$ and $m_{l+1}^\delta \to 0$ as $\delta \to 0$. By using $0 \leqslant \beta_l^\delta \leqslant \hat{\beta}_l$, we thus have $\lim_{\delta\to 0} \beta_l^\delta m_{l+1}^\delta = 0 = \beta_l m_{l+1}$. Furthermore, according to lemma 3.3, we have

$$\tilde{\gamma}_{l+1} \geqslant \langle \theta_{l+1} - \theta_l, x_l - \hat{x}\rangle = \langle m_{l+1}, x_l - \hat{x}\rangle = 0.$$

Therefore

$$\lim_{\delta\to 0} \tilde{\gamma}_{l+1}^\delta = \tilde{\gamma}_{l+1} \geqslant 0.$$

If $\tilde{\gamma}_{l+1} = 0$, we can use $0 \leqslant \beta_l^\delta \leqslant \hat{\beta}_l$ to obtain $\lim_{\delta\to 0} \beta_l^\delta \tilde{\gamma}_{l+1}^\delta = 0 = \beta_l \tilde{\gamma}_{l+1}$. If $\tilde{\gamma}_{l+1} > 0$, then

$$\frac{\alpha_l^\delta \langle g_l^\delta, m_{l+1}^\delta\rangle - 2\sigma\tilde{\gamma}_{l+1}^\delta}{\|m_{l+1}^\delta\|^2} \to -\infty \quad \text{as } \delta \to 0.$$

Consequently, by the definition of β_l^δ, we must have $\beta_l^\delta = 0$ for small $\delta > 0$, and consequently $\beta_l^\delta \tilde{\gamma}_{l+1}^\delta = 0 = \beta_l \tilde{\gamma}_{l+1}$ for small $\delta > 0$. We therefore establish (52).

Finally we show $\xi_{l+1}^\delta \to \xi_{l+1}$ and $x_{l+1}^\delta \to x_{l+1}$ as $\delta \to 0$. By the definition of ξ_{l+1}^δ and ξ_{l+1}, we can use (51) and (52) to obtain

$$\xi_{l+1}^\delta = \theta_{l+1}^\delta + \beta_l^\delta m_{l+1}^\delta \to \theta_{l+1} + \beta_l m_{l+1} = \xi_{l+1} \quad \text{as } \delta \to 0.$$

By the continuity of $\nabla\mathcal{R}^*$, we then have

$$x_{l+1}^\delta = \nabla\mathcal{R}^*\left(\xi_{l+1}^\delta\right) \to \nabla\mathcal{R}^*\left(\xi_{l+1}\right) = x_{l+1} \quad \text{as } \delta \to 0.$$

The proof is therefore complete. □

Now we are ready to show the regularization property of algorithm 1 for solving ill-posed inverse problems.

Theorem 4.5. *Let assumptions 1 and 2 hold with X being a Hilbert space and $p = 2$. Consider algorithm 1, where*

$$0 < \mu_0 < 4\sigma, \quad \mu_1 > 0, \quad \tau > \frac{1+\eta}{1-\eta}, \quad \beta := \sup_{k\geqslant 0} \hat{\beta}_k < 1.$$

Let k_δ be the output integer. Then there exists a solution x^* of (1) in $B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$ such that

$$\lim_{\delta\to 0} D_{\mathcal{R}}^{\xi_{k_\delta}^\delta}\left(x^*, x_{k_\delta}^\delta\right) = 0 \quad \text{and} \quad \lim_{\delta\to 0} \|x_{k_\delta}^\delta - x^*\| = 0.$$

If, in addition, $\mathrm{Ran}(L(x)^*) \subset \overline{\mathrm{Ran}(L(x^\dagger)^*)}$ *for all $x \in B_{2\rho}(x_0)$, then $x^* = x^\dagger$.*

Proof. Based on lemma 3.4, theorem 4.3 and lemma 4.4, the result can be proved by a standard argument [20]. For completeness we include a brief proof here.

Let $\{x_k\}$ be the sequence defined by algorithm 3 and let x^* be the solution of (1) in $B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$ determined in theorem 4.3 such that $D_{\mathcal{R}}^{\xi_k}(x^*, x_k) \to 0$ as $k \to \infty$. By setting $\Delta_k^\delta := D_{\mathcal{R}}^{\xi_k^\delta}(x^*, x_k^\delta)$ for $k \geqslant 0$, we will show $\Delta_{k_\delta}^\delta \to 0$ as $\delta \to 0$.

If there is a sequence $\{y^{\delta_l}\}$ of noisy data satisfying $\|y^{\delta_l} - y\| \leqslant \delta_l$ with $\delta_l \to 0$ such that $k_l := k_{\delta_l} \to \hat{k}$ as $l \to \infty$ for some finite integer $\hat{k}$, then

$$\|F\left(x_{\hat{k}}^{\delta_l}\right) - y^{\delta_l}\| \leqslant \tau \delta_l$$

for large l. By taking $l \to \infty$ and using lemma 4.4 and the continuity of F, we can obtain $F(x_{\hat{k}}) = y$. Thus, according to lemma 4.1 with $\hat{x} = x_{\hat{k}}$ and the property of $\mathcal{R}$, we can deduce that $x_k = x_{\hat{k}}$ for all $k \geqslant \hat{k}$. Since $x_k \to x^*$ as $k \to \infty$, we must have $x_{\hat{k}} = x^*$ and thus, by lemma 4.4 and the lower semi-continuity of $\mathcal{R}$, we can obtain

$$\limsup_{l\to\infty} \Delta_{k_l}^{\delta_l} = \mathcal{R}\left(x_{\hat{k}}\right) - \liminf_{l\to\infty} \mathcal{R}\left(x_{\hat{k}}^{\delta_l}\right) - \lim_{l\to\infty} \langle \xi_{\hat{k}}^{\delta_l}, x_{\hat{k}} - x_{\hat{k}}^{\delta_l} \rangle \leqslant \mathcal{R}\left(x_{\hat{k}}\right) - \mathcal{R}\left(x_{\hat{k}}\right) = 0$$

which shows $\Delta_{k_l}^{\delta_l} \to 0$ as $l \to \infty$.

If there is a sequence $\{y^{\delta_l}\}$ of noisy data satisfying $\|y^{\delta_l} - y\| \leqslant \delta_l$ with $\delta_l \to 0$ such that $k_l := k_{\delta_l} \to \infty$ as $l \to \infty$, then for any fixed integer k, we may use lemmas 3.4, 4.4 and the lower semi-continuity of $\mathcal{R}$ to obtain

$$\limsup_{l\to\infty} \Delta_{k_l}^{\delta_l} \leqslant \limsup_{l\to\infty} \Delta_k^{\delta_l} \leqslant \mathcal{R}(x^*) - \mathcal{R}(x_k) - \langle \xi_k, x^* - x_k \rangle = D_{\mathcal{R}}^{\xi_k}(x^*, x_k).$$

Letting $k \to \infty$ in the above equation gives $\limsup_{l\to\infty} \Delta_{k_l}^{\delta_l} \leqslant 0$ which shows again $\Delta_{k_l}^{\delta_l} \to 0$ as $l \to \infty$. □

Note that the proof of theorem 4.5 does not need X to be Hilbertian and $p = 2$ once the result in lemma 3.4 is available; the conditions that X is Hilbertian and $p = 2$ are only used to derive the formula for α_k^δ and β_k^δ in algorithm 1. Therefore, the above argument can be easily adapted to prove the regularization property of algorithm 2 as stated in the following result.

Theorem 4.6. *Let assumptions 1 and 2 hold. Consider algorithm 2, where*

$$0 < \mu_0 < 2\sigma, \quad \mu_1 > 0, \quad t > 1, \quad \tau > \frac{1+\eta}{1-\eta}, \quad \beta := \sup_{k\geqslant 0} \hat{\beta}_k < 1.$$

Let k_δ be the output integer. Then there exists a solution x^* of (1) in $B_{2\rho}(x_0) \cap \mathrm{dom}(\mathcal{R})$ such that

$$\lim_{\delta\to 0} D_{\mathcal{R}}^{\xi_{k_\delta}^\delta}\left(x^*, x_{k_\delta}^\delta\right) = 0 \quad \text{and} \quad \lim_{\delta\to 0} \|x_{k_\delta}^\delta - x^*\| = 0.$$

If, in addition, $\mathrm{Ran}(L(x)^*) \subset \overline{\mathrm{Ran}(L(x^\dagger)^*)}$ *for all* $x \in B_{2\rho}(x_0)$*, then* $x^* = x^\dagger$.

5. Numerical results

In this section we will provide various numerical experiments to demonstrate the performance of our adaptive Nesterov momentum method. We will compare our method with the Landweber type method (3) to illustrate its acceleration effect. Furthermore, we will also provide numerical results for Nesterov acceleration method (6) as comparison to numerically demonstrate further fast convergence property of our method. Note that, for Nesterov acceleration method (6), no general convergence theory is available expect for those special cases mentioned in section 1.

Example 5.1. Consider the first kind integral equation

$$(Fx)(s) := \int_0^1 \kappa(s,t)\, x(t)\, \mathrm{d}t = y(s), \quad s \in [0,1],$$

where

$$\kappa(s,t) = d\left[d^2 + (s-t)^2\right]^{-3/2}$$

with $d = 0.1$. This is a one-dimensional model problem in gravity surveying [11] which aims to recover a mass distribution $x(t)$ located at depth d from the measured vertical component of the gravity field $y(s)$ at the surface. Clearly F is a compact linear operator from $L^2[0,1]$ to itself and the problem is severely ill-posed. Assume the sought solution is

$$x^\dagger(t) = 4t(1-t) + \sin(2\pi t).$$

We calculate the exact data $y := Fx^\dagger$ and add random noise onto y to produce a noisy data y^δ satisfying $\|y^\delta - y\|_{L^2[0,1]} = \delta$ for various noise levels $\delta > 0$. We next use y^δ to reconstruct $x^\dagger$.

In order to use algorithm 1 to determine the minimal norm solution, we take $\mathcal{R}(x) = \frac{1}{2}\|x\|^2_{L^2[0,1]}$ and use the initial guess $x_0^\delta = 0$. Clearly $\mathcal{R}$ satisfies assumption 1 with $p = 2$ and $\sigma = 1/2$ and assumption 2 is satisfied with $\eta = 0$. When implementing algorithm 1 we use the following parameters

$$\tau = 1.01, \quad \mu_0 = 1.4\sigma, \quad \mu_1 = 100, \quad s = 2, \quad \hat{\beta}_k = \min\left\{0.999, \frac{k+1}{k+2}\right\}.$$

As comparisons we also execute the Landweber iteration

$$x_{k+1}^\delta = x_k^\delta - \alpha F^*\left(Fx_k^\delta - y^\delta\right) \tag{53}$$

with $x_0^\delta = 0$ and $\alpha = 1.8/\|F\|^2$, terminated by the discrepancy principle (8) with $\tau = 1.01$, and the Nesterov acceleration method

$$z_k^\delta = x_k^\delta + \frac{k-1}{k+\gamma}\left(x_k^\delta - x_{k-1}^\delta\right), \quad x_{k+1}^\delta = z_k^\delta - \alpha F^*\left(Fz_k^\delta - y^\delta\right) \tag{54}$$

with $x_0^\delta = x_{-1}^\delta = 0$, $\gamma = 3$ and $\alpha = 0.9/\|F\|^2$, terminated by the discrepancy principle (9) with $\tau = 1.01$. The convergence for Landweber iteration (53) and Nesterov acceleration method (54) are available in [6] and [27] respectively. In our implementation, all integrals over $[0,1]$ are approximated by the trapezoidal rule by partitioning $[0,1]$ into 1000 subintervals of equal length.

In table 1 we report the computational results by our algorithm 1, Landweber iteration (53) and Nesterov acceleration method (54), including the required number of iterations, the consumed CPU time, and the corresponding relative errors. The results indicate that these three methods can produce comparable approximate solutions in terms of accuracy. However, our algorithm 1 and Nesterov acceleration method (54) clearly demonstrate superior performance over the Landweber method (53) by significantly reducing the number of iterations and the CPU running time. To visualize the acceleration effect, in figure 1 we plot on the left the relative error $\|x_k^\delta - x^\dagger\|_{L^2}/\|x^\dagger\|_{L^2}$ versus k for these three methods using noisy data with $\delta = 0.001$, until the iteration is terminated by the discrepancy principle; the right figure plots the output momentum coefficients β_k^δ by our algorithm 1.

Table 1. Computational results for Example 5.1.

δ	method	iterations	time (s)	relative error
0.1	Landweber	69	0.0205	1.5804×10^{-2}
	Nesterov	29	0.0082	1.7047×10^{-2}
	Adaptive Nesterov	29	0.0105	1.6461×10^{-2}
0.01	Landweber	965	0.1579	7.2476×10^{-3}
	Nesterov	132	0.0177	7.1879×10^{-3}
	Adaptive Nesterov	90	0.0236	7.1184×10^{-3}
0.001	Landweber	15 561	2.4608	3.5064×10^{-3}
	Nesterov	558	0.0972	3.4416×10^{-3}
	Adaptive Nesterov	475	0.0958	3.5019×10^{-3}
0.0001	Landweber	272 037	44.623	1.6609×10^{-3}
	Nesterov	2358	0.3987	1.6371×10^{-3}
	Adaptive Nesterov	2143	0.3575	1.6597×10^{-3}

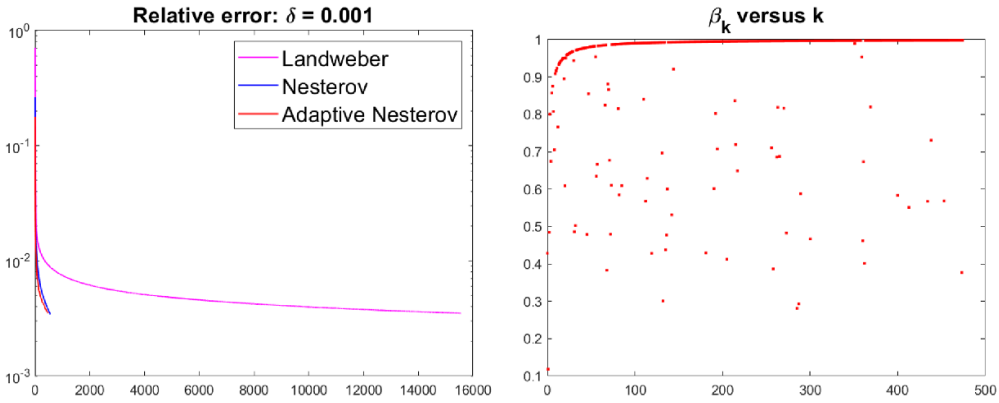

Figure 1. Example 5.1: relative errors versus the number of iterations and the plot of momentum coefficients.

In the following we will provide further numerical experiments in which the sought solutions admit special features requiring to use non-quadratic functions $\mathcal{R}$ for reconstructions. We will consider the Landweber type method (3) terminated by (4) with the step-size given by (5). Further, we will consider Nesterov acceleration method (6) terminated by

$$\|F\left(z_{k_\delta}^\delta\right) - y^\delta\| \leqslant \tau\delta < \|F\left(z_k^\delta\right) - y^\delta\|, \quad 0 \leqslant k < k_\delta \tag{55}$$

with α_k^δ given by

$$\alpha_k^\delta = \min\left\{\frac{\mu_0\|F\left(z_k^\delta\right) - y^\delta\|^{(p-1)s}}{\|L\left(z_k^\delta\right)^* J_s^Y\left(F\left(z_k^\delta\right) - y^\delta\right)\|^p}, \mu_1\|F\left(z_k^\delta\right) - y^\delta\|^{p-s}\right\}, \tag{56}$$

we will then use $x_{k_\delta}^\delta$ output by (7) as an approximate solution. The equations (55) and (56) are different from (4) and (5) in that they use z_k^δ instead of x_k^δ. This adjustment can save the

computational work of (6) by avoiding the calculation of x_k^δ and $\|F(x_k^\delta) - y^\delta\|$ during the iteration process. We should emphasize that all the following computations on the method (6) are empirical, since there is no convergence theory available for (6) terminated by (55) with the step-size chosen by (56).

Example 5.2. We next consider the first kind linear integral equation

$$(Fx)(s) := \int_0^1 \kappa(s,t)\, x(t)\, \mathrm{d}t = y(s), \quad s \in [0,1],$$

where $\kappa(s,t) = 4e^{-(s-t)^2/0.01}$, and assume the sought solution is nonnegative. Clearly F is a compact linear operator from $L^2[0,1]$ to itself. In order to find the unique nonnegative solution $x^\dagger$ with minimal norm, we take $\mathcal{R}(x) = \frac{1}{2}\|x\|^2_{L^2[0,1]} + \iota_C(x)$, where $C := \{x \in L^2[0,1] : x \geqslant 0 \text{ a.e. on } [0,1]\}$. This $\mathcal{R}$ satisfies assumption 1 with $p = 2$ and $\sigma = 1/2$. Note that the resolution of the minimization problem

$$x = \arg\min_{z \in L^2[0,1]} \{\mathcal{R}(x) - \langle \xi, z\rangle\}$$

is required at each iteration step in the implementation of the Landweber type method (3), the Nesterov acceleration method (6), and our algorithm 1; the solution is given by

$$x = \arg\min_{z \in C} \left\{ \frac{1}{2}\|z\|^2_{L^2[0,1]} - \langle \xi, z\rangle \right\} = \max\{\xi, 0\}.$$

In our numerical computation, we assume the sought solution is

$$x^\dagger(t) = \max\{20t(t-0.2)(0.75-t), 0\}.$$

Let $y := Fx^\dagger$ be the exact data. We add random noise on y to produce a noisy data y^δ satisfying $\|y^\delta - y\|_{L^2[0,1]} = \delta$ for various noise levels $\delta > 0$. To carry out the computation, all integrals over $[0,1]$ are approximated by the trapezoidal rule based on the 1000 nodes partitioning $[0,1]$ into subintervals of equal length. We execute the Landweber type method (3), Nesterov acceleration method (6) and our algorithm 1, and all these methods are terminated by the discrepancy principle with $\tau = 1.01$. For the Landweber type method and Nesterov acceleration method, we use $\mu_0 = 0.99(2 - 2/\tau)$ and $\mu_1 = 100$ in their step size choices; we also use $\gamma = 5$ in the method (6). For our algorithm 1 we use the parameters

$$\mu_0 = 0.8, \quad \mu_1 = 100, \quad s = 2, \quad \hat{\beta}_k = \min\{0.98, (k+1)/(k+2)\}.$$

The computational results are reported in table 2 which includes the required number of iterations, the consumed CPU time, and the corresponding relative errors. These results indicate that Nesterov acceleration method (6) and our algorithm 1 clearly have the acceleration effect over the Landweber-type method. Our algorithm 1 can even converge faster than Nesterov acceleration method. Furthermore, our algorithm 1 can produce more accurate results than the Landweber-type iteration. In order to visualize how the iterates approach the sought solution, in figure 2 we plot the relative error $\|x_k^\delta - x^\dagger\|_{L^2}/\|x^\dagger\|_{L^2}$ versus k for the three methods: the left figure plots the relative error, using noisy data with $\delta = 0.0001$, until the iteration is terminated by the discrepancy principle; the right figure plots the momentum coefficient β_k^δ determined by algorithm 1 during the computation.

Table 2. Computational results for Example 5.2.

δ	method	iterations	time (s)	relative error
0.1	Landweber	53	0.0091	9.4104×10^{-2}
	Nesterov	23	0.0052	7.2489×10^{-2}
	Adaptive Nesterov	6	0.0015	7.4869×10^{-2}
0.01	Landweber	184	0.0331	3.3619×10^{-2}
	Nesterov	50	0.0128	2.7179×10^{-2}
	Adaptive Nesterov	14	0.0029	1.5487×10^{-2}
0.001	Landweber	680	0.1135	1.0286×10^{-2}
	Nesterov	112	0.0213	8.9134×10^{-3}
	Adaptive Nesterov	39	0.0079	5.0230×10^{-3}
0.0001	Landweber	4808	0.6892	3.6475×10^{-3}
	Nesterov	321	0.0563	2.9460×10^{-3}
	Adaptive Nesterov	148	0.0255	2.4844×10^{-3}
0.000 01	Landweber	42 944	6.8886	1.6496×10^{-3}
	Nesterov	806	0.1263	1.6244×10^{-3}
	Adaptive Nesterov	307	0.0588	1.0969×10^{-3}

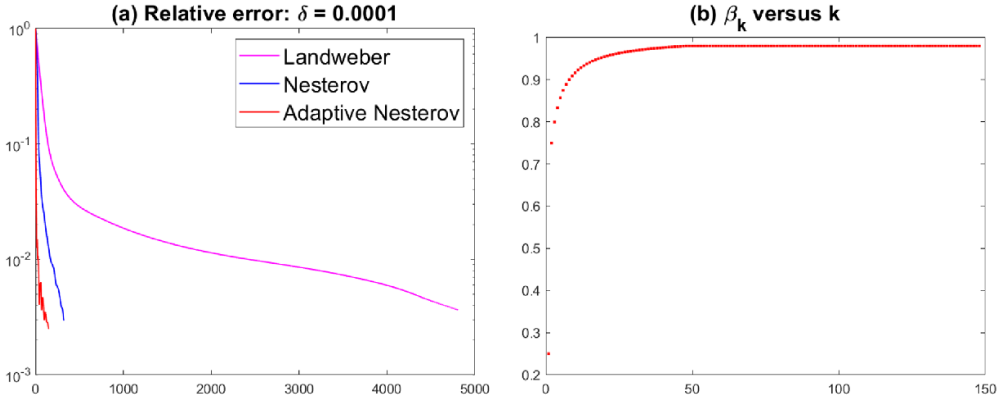

Figure 2. Example 5.2: relative errors versus the number of iterations and the plot of momentum coefficients.

Example 5.3. In this example, we consider the standard 2D fan-beam x-ray tomography using simulated tomographic data to illustrate the performance of algorithm 1. This modality involves reconstructing a body slice by collecting x-ray attenuation data, mathematically expressed as finding a compactly supported function from its Radon transform.

We discretize the sought image on a 256×256 pixel grid and identify it by a long vector in $\mathbb{R}^N$ with $N = 256 \times 256 = 65\,536$ by stacking all its columns. For reconstruction we use tomographic data with $p = 60$ projections and 367 x-ray lines per projection. By using the function `fanbeamtomo` from the MATLAB package AIR TOOLS [12] to discretize the problem, we can obtain an ill-conditioned linear system $Ax = y$, where A is a coefficient matrix of size $22\,020 \times 65\,536$. The true image $x^\dagger$ in our experiment is the modified Shepp–Logan phantom. We calculate $y := Ax^\dagger$ and add Gaussian noise on y to generate a noisy data y^δ with relative

Table 3. Computational results for Example 5.3.

$\delta_{\rm rel}$	method	iterations	time (s)	relative error
5%	Landweber	101	9.9940	1.8538×10^{-1}
	Nesterov	34	3.2924	1.6614×10^{-1}
	Adaptive Nesterov	25	2.5897	1.5015×10^{-1}
1%	Landweber	444	44.005	5.9528×10^{-2}
	Nesterov	79	7.1505	5.1758×10^{-2}
	Adaptive Nesterov	59	5.6087	3.7336×10^{-2}
0.5%	Landweber	778	78.410	3.1650×10^{-2}
	Nesterov	104	9.1250	2.4194×10^{-2}
	Adaptive Nesterov	103	9.6334	1.7636×10^{-2}
0.1%	Landweber	3254	366.63	6.2999×10^{-3}
	Nesterov	326	27.146	2.8395×10^{-3}
	Adaptive Nesterov	227	21.951	3.3969×10^{-3}

noise level $\delta_{\rm rel} = \|y^\delta - y\|_2/\|y\|_2$ so that the noise level is $\delta = \delta_{\rm rel}\|y\|_2$. In order to capture the feature of the sought image, we take

$$\mathcal{R}(x) = \frac{1}{2\kappa}\|x\|_{\rm F}^2 + |x|_{\rm TV}, \quad \forall x \in \mathbb{R}^{256\times 256}$$

with $\kappa = 1$, where $\|x\|_{\rm F}$ is the Frobenius norm and $|x|_{\rm TV}$ denotes the isotropic total variation of x, see [15, 37]. Clearly $\mathcal{R}$ satisfies assumption 1 with $p = 2$ and $\sigma = 1/(2\kappa)$. When using algorithm 1 to reconstruct the image, we use the initial guess $\xi_0 = 0$ and the following parameters

$$\tau = 1.05, \quad \mu_0 = 0.6/\kappa, \quad \mu_1 = 100, \quad s = 2, \quad \hat\beta_k = \min\left\{0.98, \frac{k+1}{k+3}\right\}.$$

As comparison, we also carry out the computation by the Landweber type method (3) and the Nesterov acceleration method (6) terminated by the discrepancy principle with $\tau = 1.05$ using $\mu_0 = 0.99(2 - 2/\tau)/\kappa$ and $\mu_1 = 100$ in the step-size; $\gamma = 5$ is used in the method (6). During the computation, updating x_k^δ from ξ_k^δ requires to solving the total variation denoising problem

$$x_k^\delta = \arg\min_{x\in\mathbb{R}^{256\times 256}}\left\{\frac{1}{2\kappa}\|x - \kappa\xi_k^\delta\|_{\rm F}^2 + |x|_{\rm TV}\right\}$$

which is solved approximately by the primal-dual hybrid gradient (PDHG) method [37] after 70 iterations.

The computational results by these methods are reported in table 3, including the number of iterations k_δ, the CPU running time, and the relative errors $\|x_{k_\delta}^\delta - x^\dagger\|_2/\|x^\dagger\|_2$, using noisy data with various relative noise level $\delta_{\rm rel} > 0$. Table 3 shows that algorithm 1 leads to a considerable reduction on the number of iterations and the amount of computational time, which demonstrates the striking acceleration effect of our adaptive Nesterov momentum method. To visualize the performance, we plot in figure 3 the true image and the reconstruction results by the above three methods using noisy data with relative noise level $\delta_{\rm rel} = 1\%$ together with the relative errors $\|x_k^\delta - x^\dagger\|_2/\|x^\dagger\|_2$ and the output momentum coefficient β_k^δ.

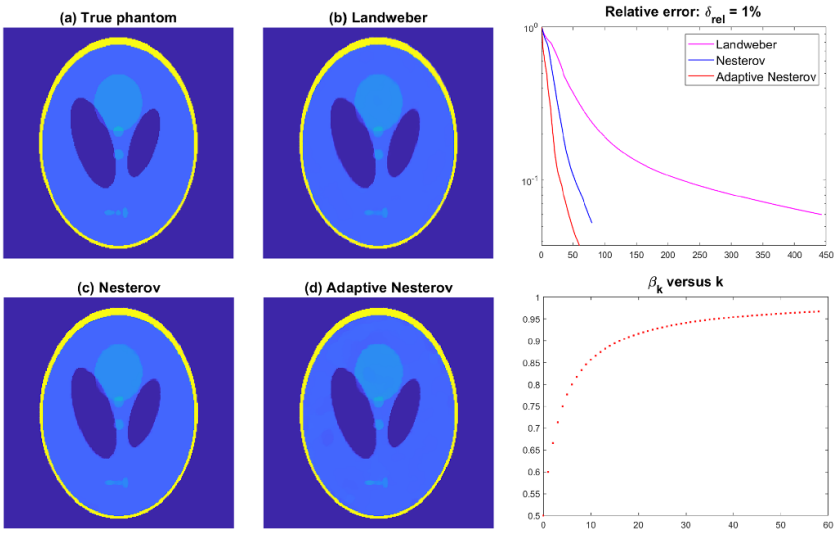

Figure 3. Example 5.3: reconstruction results, relative errors versus the number of iterations, and the plot of momentum coefficients.

Example 5.4. In this example we consider determining the coefficient c in the elliptic boundary value problem

$$-\triangle u + cu = f \text{ in } \Omega, \qquad u = g \text{ on } \partial\Omega \tag{57}$$

from an $L^2(\Omega)$-measurement of the state u, where $\Omega := [0,1]^2 \subset \mathbb{R}^2$, $f \in H^{-1}(\Omega)$ and $g \in H^{1/2}(\Omega)$; see [6]. We assume the sought coefficient $c^\dagger$ is in $L^2(\Omega)$. This nonlinear inverse problem reduces to solving $F(c) = u$ with $F : L^2(\Omega) \to L^2(\Omega)$ defined by $F(c) := u(c)$, where $u(c) \in H^1(\Omega) \subset L^2(\Omega)$ is the unique solution of (57). This nonlinear operator F is well defined on

$$\mathcal{D} := \left\{ c \in L^2(\Omega) : \|c - \hat{c}\|_{L^2(\Omega)} \leqslant \varepsilon_0 \text{ for some } \hat{c} \geqslant 0, \text{ a.e.} \right\}$$

for some positive constant $\varepsilon_0 > 0$. It is known that the operator F is weakly closed and Fréchet differentiable, the Fréchet derivative of F and its adjoint are given by

$$F'(c)h = -A(c)^{-1}(hu(c)) \quad \text{and} \quad F'(c)^* w = -u(c)A(c)^{-1} w$$

for $c \in \mathcal{D}$ and $h, w \in L^2(\Omega)$, where $A(c) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by $A(c)u = -\triangle u + cu$ which is an isomorphism uniformly in the ball $B_{2\rho}(c^\dagger)$ for small $\rho > 0$. Moreover, F satisfies assumption 2 (iii) with $L(x) = F'(x)$ for a number $\eta > 0$ which can be very small if $\rho > 0$ is sufficiently small (see [6]). In our numerical computation below we will take $\eta = 0.01$.

We consider the situation that the sought solution $c^\dagger$ is piece-wise constant and assume it is given by

$$\begin{aligned} c^\dagger(x,y) = &\max\left\{0, \operatorname{sign}\left(0.13^2 - (x-0.25)^2 - (y-0.70)^2\right)\right\} \\ &+ 0.5\max\left\{0, \operatorname{sign}\left(0.07^2 - (x-0.35)^2/4 - (y-0.3)^2\right)\right\} \\ &+ 0.6\max\left\{0, \operatorname{sign}\left(0.15 - |x-0.70| - |y-0.5|\right)\right\}; \end{aligned}$$

Table 4. Computational results for Example 5.4.

δ	method	iterations	time (s)	$\|x_{n_\delta}^\delta - x^\dagger\|_{L^2}$
0.004	Landweber	18	3.5713	2.6003×10^{-1}
	Nesterov	20	4.1421	2.4669×10^{-1}
	Adaptive Nesterov	14	2.8167	2.6220×10^{-1}
0.001	Landweber	179	30.644	1.4552×10^{-1}
	Nesterov	72	13.931	1.1968×10^{-1}
	Adaptive Nesterov	86	15.212	1.1344×10^{-1}
0.0004	Landweber	566	100.54	1.0892×10^{-1}
	Nesterov	139	26.598	9.7177×10^{-2}
	Adaptive Nesterov	113	18.872	8.7846×10^{-2}
0.0001	Landweber	2329	542.18	8.0650×10^{-2}
	Nesterov	330	61.996	7.3898×10^{-2}
	Adaptive Nesterov	240	41.493	6.6211×10^{-2}
0.000 04	Landweber	10 357	2366.8	6.6261×10^{-2}
	Nesterov	722	136.06	5.7057×10^{-2}
	Adaptive Nesterov	384	70.515	6.1092×10^{-2}

the graph of this function is plotted in figure 4(a). In order to find such a coefficient, we use

$$\mathcal{R}(c) := \frac{1}{2\kappa}\|c\|^2_{L^2(\Omega)} + \int_\Omega |Dc|$$

with $\kappa = 10$, where $\int_\Omega |Dc|$ denotes the total variation of c on Ω. Clearly this $\mathcal{R}$ satisfies assumption 1 with $p = 2$ and $\sigma = 1/(2\kappa)$. Assuming $u(c^\dagger) = x + y$, we add random noise to produce noisy data u^δ satisfying $\|u^\delta - u(c^\dagger)\|_{L^2(\Omega)} = \delta$ with various noise level $\delta > 0$. We will use u^δ to reconstruct $c^\dagger$.

We carry out the computation by our algorithm 1 using the initial guess $\xi_0 = 0$ and the parameters

$$\tau = 1.05, \quad \mu_0 = 1, \quad \mu_1 = 600, \quad s = 2, \quad \hat{\beta}_k = \min\{0.98, (k+1)/(k+2)\}.$$

We also execute the Landweber type method (3) and Nesterov acceleration method (6) terminated by the discrepancy principle with $\tau = 1.05$; we use the same initial guess ξ_0 with $\mu_0 = 1.8(1-\eta-(1+\eta)/\tau)/\kappa$ and $\mu_1 = 600$ in the step-size choices. In order to carry out the computation, we divide Ω into 128×128 small squares of equal size and solve all partial differential equations approximately by a multigrid method [8] via finite difference discretization. Furthermore, updating c_n^δ from ξ_n^δ at each iteration step is equivalent to a total variation denoising problem which is solved by the PDHG method after 200 iterations.

In table 4 we report the computational results by algorithm 1, the Landweber type method (3), and Nesterov acceleration method (6), including the required number of iterations k_δ, the CPU running time and the absolute errors $\|c_{k_\delta}^\delta - c^\dagger\|_{L^2(\Omega)}$, using noisy data for various noise level $\delta > 0$. In figure 4 we plot the reconstructed results and the absolute errors of these three methods using noisy data with noise level $\delta = 0.0001$, together with the momentum

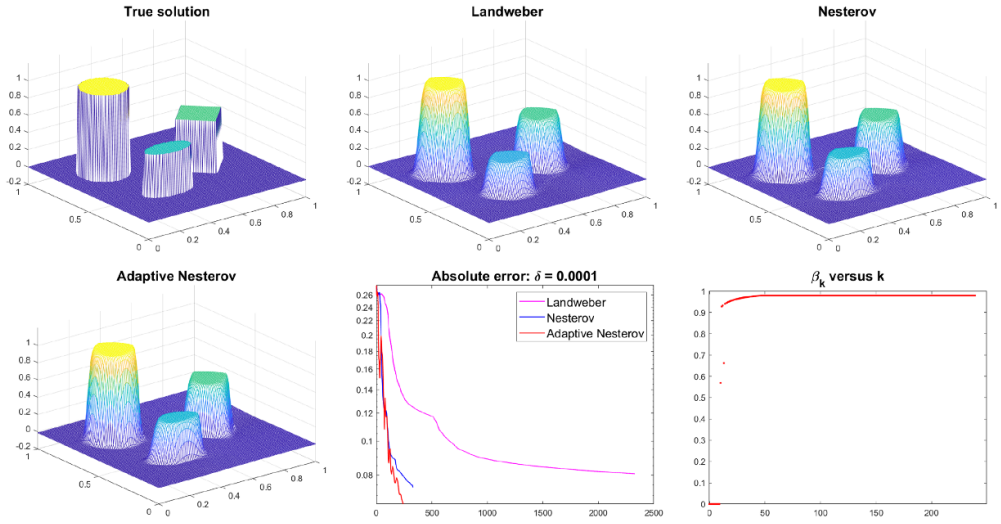

Figure 4. Example 5.4: reconstruction results, relative errors versus the number of iterations, and the plot of momentum coefficients.

coefficients β_k output by algorithm 1. These results clearly demonstrate the acceleration effect of algorithm 1, when compared with the Landweber-type method.

Example 5.5. In this last example we consider the situation that the sought solution lies in a Banach space and algorithm 2 needs to be used for reconstruction. We consider the linear integral equation

$$(Fx)(u) := \int_0^1 \kappa(u,v)x(v)\,dv = y(u), \quad u \in [0,1],$$

where $\kappa(u,v) = 4e^{-(u-v)^2/0.01}$, and assume the sought solution $x^\dagger$ is a probability density function, i.e. $x^\dagger \geqslant 0$ a.e. on $[0,1]$ and $\int_0^1 x^\dagger = 1$. Clearly F is a compact linear operator from $L^1[0,1]$ to $L^2[0,1]$. We determine such a solution by using

$$\mathcal{R}(x) = f(x) + \iota_\Delta(x),$$

where ι_Δ denotes the indicator function of

$$\Delta := \left\{ x \in L^1_+[0,1] : \int_0^1 x^\dagger = 1 \right\}$$

and f denotes the negative of the Boltzmann-Shannon entropy, i.e.

$$f(x) = \begin{cases} \int_0^1 x\log x, & \text{if } x \in L^1_+[0,1] \text{ and } x\log x \in L^1[0,1], \\ +\infty, & \text{otherwise.} \end{cases}$$

Here $L^1_+[0,1] := \{x \in L^1[0,1] : x \geqslant 0 \text{ a.e. on } [0,1]\}$. It is known that $\mathcal{R}$ satisfies assumption 1 with $p = 2$ and $\sigma = 1/2$, moreover, for any $\xi \in L^\infty[0,1]$ there holds

Table 5. Computational results for Example 5.5.

δ	method	iterations	time (s)	relative error
0.1	Landweber	369	0.1093	8.0632×10^{-2}
	Nesterov	76	0.0296	6.6576×10^{-2}
	Adaptive Nesterov	45	0.0190	7.4328×10^{-2}
0.01	Landweber	1557	0.4253	2.8430×10^{-2}
	Nesterov	189	0.0680	2.7588×10^{-2}
	Adaptive Nesterov	172	0.0557	2.4048×10^{-2}
0.001	Landweber	11 453	2.9894	1.3897×10^{-2}
	Nesterov	497	0.1388	1.5214×10^{-2}
	Adaptive Nesterov	344	0.1007	9.4030×10^{-3}
0.0001	Landweber	142 426	47.875	4.4288×10^{-3}
	Nesterov	1593	0.4632	5.0806×10^{-3}
	Adaptive Nesterov	761	0.2065	4.2279×10^{-3}

$$\arg\min_{x\in L^1[0,1]} \{\mathcal{R}(x) - \langle \xi, x\rangle\} = e^{\xi} / \int_0^1 e^{\xi};$$

see [16, 19] for instance.

For numerical simulations, we assume the sought solution is

$$x^\dagger(v) := c\left(e^{-60(v-0.3)^2} + 0.3e^{-40(v-0.8)^2}\right),$$

where $c > 0$ is a constant to ensure that $\int_0^1 x^\dagger(v)dv = 1$ so that $x^\dagger$ is a probability density function. Let $y := Fx^\dagger$ be the exact data, we add random noise on y to produce a noisy data y^δ satisfying $\|y^\delta - y\|_{L^2[0,1]} = \delta$ for various noise levels $\delta > 0$. We execute the Landweber type method (3), Nesterov acceleration method (6) and our algorithm 2; all these methods are terminated by the discrepancy principle with $\tau = 1.01$. For the methods (3) and (6), we use $\mu_0 = 0.99(2 - 2/\tau)$ and $\mu_1 = 100$ in their the step size choices; we also use $\gamma = 5$ in (6). For our algorithm 2 we use the parameters

$$\mu_0 = 0.875, \quad \mu_1 = 100, \quad t = 5, \quad s = 2, \quad \hat{\beta}_k = \min\{0.99, (k+1)/(k+2)\}.$$

The computational results are reported in table 5 which clearly indicates that Nesterov acceleration method (6) and our algorithm 2 have the acceleration effect over the Landweber-type method (3). Furthermore, our algorithm 2 can produce more accurate results than the Landweber-type method. To visualize the performance of these methods, in figure 5 we plot their relative errors $\|x_k^\delta - x^\dagger\|_{L^1} / \|x^\dagger\|_{L^1}$ using noisy data with $\delta = 0.01$ together with the momentum coefficient β_k^δ produced by algorithm 2.

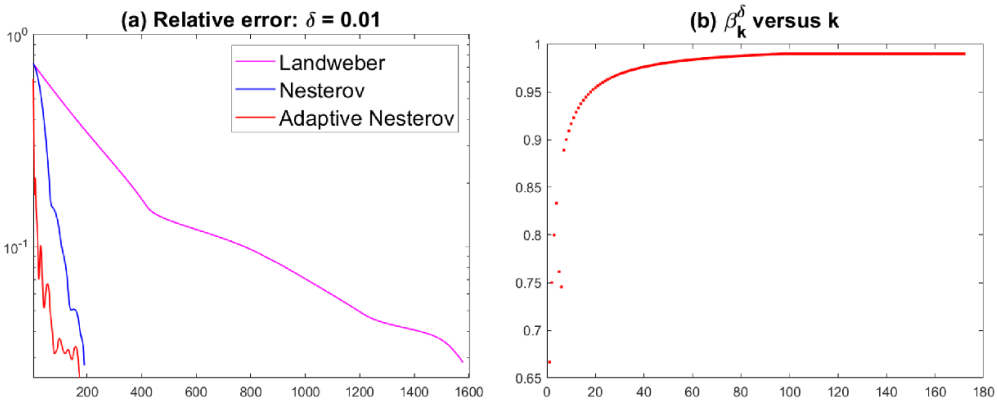

Figure 5. Example 5.5: relative errors versus the number of iterations, and the plot of momentum coefficients.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iD

Qinian Jin https://orcid.org/0000-0001-9283-9791

References

[1] Attouch H and Peypouquet J 2016 The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $(1/k^2)$ *SIAM J. Optim.* **26** 1824–34
[2] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202
[3] Blaschke B, Neubauer A and Scherzer O 1997 On convergence rates for the iteratively regularized Gauss-Newton method *IMA J. Numer. Anal.* **17** 421–36
[4] Bot R and Hein T 2012 Iterative regularization with a general penalty term—theory and applications to L^1 and TV regularization *Inverse Problems* **28** 104010
[5] Cioranescu I 1999 *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems* (Kluwer)
[6] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Kluwer)
[7] Frick K and Scherzer O 2010 Regularization of ill-posed linear equations by the non-stationary augmented Lagrangian method *J. Integral Equ. Appl.* **22** 217–57
[8] Hackbusch W 2016 *Iterative Solution of Large Sparse Systems of Equations* (*Applied Mathematical Sciences* vol 95) 2nd edn (Springer)
[9] Hanke M 1997 A regularizing Levenberg-Marquardt scheme with applications to inverse groundwater filtration problems *Inverse Problems* **13** 79–95
[10] 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
[11] Hansen P C 2007 Regularization tools version 4.0 for Matlab 7.3 *Numer. Algorithms* **46** 189–94
[12] Hansen P C and Saxild-Hansen M 2012 AIR tools—a MATLAB package of algebraic iterative reconstruction methods *J. Comput. Appl. Math.* **236** 2167–78
[13] Hubmer S and Ramlau R 2017 Convergence analysis of a two-point gradient method for nonlinear ill-posed problems *Inverse Problems* **33** 095004

[14] Hubmer S and Ramlau R 2018 Nesterov's accelerated gradient method for nonlinear ill-posed problems with a locally convex residual functional *Inverse Problems* **34** 095003
[15] Jin Q 2016 Landweber-Kaczmarz method in Banach spaces with inexact inner solvers *Inverse Problems* **32** 104005
[16] Jin Q 2022 Convergence rates of a dual gradient method for constrained linear ill-posed problems *Numer. Math.* **151** 841–71
[17] Jin Q and Huang Q 2024 An adaptive heavy ball method for ill-posed inverse problems *SIAM J. Imaging Sci.* **17** 2212–41
[18] Jin Q and Lu X 2014 A fast nonstationary iterative method with convex penalty for inverse problems in Hilbert spaces *Inverse Problems* **30** 045012
[19] Jin Q, Lu X and Zhang L 2023 Stochastic mirror descent methods for linear ill-posed problems in Banach spaces *Inverse Problems* **39** 065010
[20] Jin Q and Wang W 2013 Landweber iteration of Kaczmarz type with general non-smooth convex penalty functionals *Inverse Problems* **29** 085011
[21] Jin Q and Yang H 2016 Levenberg-Marquardt method in Banach spaces with general convex regularization terms *Numer. Math.* **133** 655–84
[22] Jin Q and Zhong M 2013 On the iteratively regularized Gauss–Newton method in Banach spaces with applications to parameter identification problems *Numer. Math.* **124** 647–83
[23] Jin Q and Zhong M 2014 Nonstationary iterated Tikhonov regularization in Banach spaces with uniformly convex penalty terms *Numer. Math.* **127** 485–513
[24] Kaltenbacher B, Neubauer A and Scherzer O 2008 *Iterative Regularization Methods for Nonlinear Ill-Posed Problems* (de Gruyter)
[25] Kaltenbacher B, Schöpfer F and Schuster T 2009 Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems *Inverse Problems* **25** 065003
[26] Kindermann S 2021 Optimal-order convergence of Nesterov acceleration for linear ill-posed problems *Inverse Problems* **37** 065002
[27] Liu D, Huang Q and Jin Q 2025 A revisit on Nesterov acceleration for linear ill-posed problems *J. Complexity* **87** 101920
[28] Louis A K 1989 *Inverse und Schlecht Gestellte Probleme* (*Teubner Studienbücher Mathematik*) (Vieweg+Teubner Verlag)
[29] Nesterov Y 1983 A method of solving a convex programming problem with convergence rate $O(1/k^2)$ *Sov. Math. Dokl.* **27** 372–6
[30] Neubauer A 2017 On Nesterov acceleration for Landweber iteration of linear ill-posed problems *J. Inverse Ill-Posed Probl.* **25** 381–90
[31] Rockafellar R T 1970 *Convex Analysis* (Princeton University Press)
[32] Scherzer O, Grasmair M, Grossauer H, Haltmeier M and Lenzen F 2008 *Variational Methods in Imaging* (*Applied Mathematical Sciences*) (Springer)
[33] Schöpfer F, Louis A K and Schuster T 2006 Nonlinear iterative methods for linear ill-posed problems in Banach spaces *Inverse Problems* **22** 311–29
[34] Schuster T, Kaltenbacher B, Hofmann B and Kazimierski K S 2012 *Regularization Methods in Banach Spaces* (*Radon Series on Computational and Applied Mathematics* vol 10) (Walter de Gruyter)
[35] Zălinscu C 2002 *Convex Analysis in General Vector Spaces* (World Scientific)
[36] Zhong M, Wang W and Jin Q 2019 Regularization of inverse problems by two-point gradient methods in Banach spaces *Numer. Math.* **143** 713–47
[37] Zhu M and Chan T F 2008 An efficient primal-dual hybrid gradient algorithm for total variation image restoration *CAM Report* 08–34 (UCLA)