Dual gradient method for ill-posed problems using multiple repeated measurement data

Qinian Jin$^1$,∗ and Wei Wang$^2$

$^1$ Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia
$^2$ College of Data Science, Jiaxing University, Jiaxing 314001, Zhejiang Province, People’s Republic of China

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

Received 25 November 2022; revised 12 May 2023
Accepted for publication 12 June 2023
Published 23 June 2023

Abstract
We consider determining $\mathcal{R}$-minimizing solutions of linear ill-posed problems $Ax = y$, where $A : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator from a Banach space $\mathcal{X}$ to a Hilbert space $\mathcal{Y}$ and $\mathcal{R} : \mathcal{X} \to [0, \infty]$ is a proper strongly convex penalty function. Assuming that multiple repeated independent identically distributed unbiased data of $y$ are available, we consider a dual gradient method to reconstruct the $\mathcal{R}$-minimizing solution using the average of these data. By terminating the method by either an a priori stopping rule or a statistical variant of the discrepancy principle, we provide the convergence analysis and derive convergence rates when the sought solution satisfies certain variational source conditions. Various numerical results are reported to test the performance of the method.

Keywords: dual gradient method, ill-posed problems, multiple repeated measurement data, convergence, convergence rates

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

∗ Author to whom any correspondence should be addressed.

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
1. Introduction

Inverse problems arise from many practical applications in science and engineering when one aims at determining unobservable causes from observed effects. Due to their significance in wide range of applications, inverse problems have received tremendous attention in the literature.

In this paper we will consider linear inverse problems of the form

\[ Ax = y, \]  

(1.1)

where \( A : \mathcal{X} \to \mathcal{Y} \) is a bounded linear operator from a Banach space \( \mathcal{X} \) to a Hilbert space \( \mathcal{Y} \). Throughout the paper we will assume that \( y \in \text{Ran}(A) \), the range of \( A \), which guarantees (1.1) has a solution. However, (1.1) may have many solutions. In order to find a solution with desired feature, by incorporating a priori available information we take a proper, lower semi-continuous, strongly convex function \( R : \mathcal{X} \to [0, \infty] \) and search for a solution \( x^\dagger \) of (1.1) satisfying

\[ R(x^\dagger) = \min \{ R(x) : x \in \mathcal{X} \text{ and } Ax = y \} \]  

(1.2)

which is called the \( R \)-minimizing solution of (1.1).

In practical applications, the exact data \( y \) is generally unknown. Very often, we only have corrupted measurement data at hand. Due to the inherent ill-posedness property of inverse problems, the \( R \)-minimizing solution of (1.1) may not depend continuously on the data. How to use the noisy data to reconstruct the sought solution therefore becomes a central topic in computational inverse problems.

To conquer ill-posedness, many iterative regularization methods have been developed to solve inverse problems using noisy data; see [5, 8, 10, 28, 29, 38] for instance. The performance of these methods depends crucially on a proper termination of the iterations. In case an accurate upper bound of noise level is available, various a posteriori rules, including the prominent discrepancy principle, have been proposed to choose the stopping index of iteration. In case the information on noise level is difficult to infer, one has to resort to the so-called heuristic rules [12, 22, 23, 30, 31] which use only the noisy data to select a termination index. Although Bakushiski’s veto [1] states that heuristic rules can not lead to convergence in the sense of worst case scenario for any regularisation method, such rules often work as well as, or even better than, the rules using information on noise level. The drawback of a heuristic rule is that it chooses a stopping index of iteration based on minimizing a function over \( \mathbb{N} := \{1, 2, \ldots\} \) for which the existence of a minimizer can not be guaranteed unless the noisy data satisfies certain restrictive conditions.

In this paper we assume that the data can be measured multiple times repeatedly and the sought solution does not change during the measurement. This is a common practice in applications where the experiments are set up to allow acquiring multiple observation data by repeating the experiments. Therefore, instead of the unknown exact data \( y \), we are given multiple unbiased measurements \( y_1, y_2, \ldots \) of \( y \) which are independent identically distributed \( \mathcal{Y} \)-valued random variables. Based on these multiple measurements, for each \( n \geq 1 \) we use

\[ \bar{y}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} y_i \]  

(1.3)

as an estimator of \( y \). We then use \( \bar{y}^{(n)} \) to construct regularized solutions and consider their behavior as \( n \to \infty \).

Using the average of multiple measurements to decrease the data error is a standard noise reduction technique in engineering which is called the signal averaging method [15, 34, 41].
Assume \((y_i)\) are \(\mathcal{Y}\)-valued random variables defined on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) with \(\mathbb{E}\) denoting the expectation. If \((y_i)\) are unbiased, independent, identically distributed estimators of \(y\) with a finite variance \(\sigma^2 := \mathbb{E}[\|y_1 - y\|^2]\), we have
\[
\mathbb{E}\left[\|\bar{y}^{(n)} - y\|^2\right] = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E}[\|y_i - y, y_j - y\|] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[\|y_i - y\|^2] = \sigma^2/n
\]
which demonstrates that using \(\bar{y}^{(n)}\) as an estimator of \(y\) can reduce the variance by a factor of \(n\). Therefore, it is reasonable to expect more accurate approximate solution can be constructed using the averaged data \(\bar{y}^{(n)}\) than using an individual data \(y_i\).

Recently some research work has been conducted to analyze linear regularization methods for linear ill-posed problems in Hilbert spaces using the average of multiple measurement data, see [13, 14, 21]. In particular a general class of linear spectral regularization methods have been considered in [13] without using any knowledge of the noise distribution, the convergence and rates of convergence of the regularized solutions, as the number of measurements increases, have been provided under either an a priori parameter choice rule or the discrepancy principle. The results obtained in [13] can be applied mainly for (1.2) with \(\mathcal{X}\) being a Hilbert space and \(\mathcal{R}(x) = \|x\|^2/2\). In many applications, however, the sought solution may sit in a Banach space instead of a Hilbert space, and the sought solution may have a priori known special features, such as nonnegativity, sparsity and piecewise constancy. Therefore, it is necessary to develop regularization methods using multiple repeated measurement data for finding a \(\mathcal{R}\)-minimizing solution of (1.1) in a general Banach space \(\mathcal{X}\) with a general strongly convex penalty function \(\mathcal{R}\).

In this paper we will consider a dual gradient method for solving (1.2) using multiple repeated measurement data. When the exact data \(y\) is available, this method can be derived by applying the gradient method to the dual problem of (1.2) and it takes the form (see [24])
\[
\begin{align*}
x_t &= \arg \min_{x \in \mathcal{X}} \left\{ \mathcal{R}(x) - \langle \lambda_t, Ax \rangle \right\}, \\
\lambda_{t+1} &= \lambda_t - \gamma (Ax_t - y) \\
&= (1.4)
\end{align*}
\]
with the initial guess \(\lambda_0 = 0\), where \(A^* : \mathcal{Y} \to \mathcal{X}^*\) denotes the adjoint of \(A\) and \(\gamma > 0\) is a step-size. This method actually belongs to the class of Uzawa methods, see [3, 39, 40]. This method also has connection to the mirror descent method, see [5, 24, 29, 35]. In case the exact data \(y\) is unavailable and we are given multiple repeated random measurement data \(y_1, y_2, \ldots\), we may define \(\bar{y}^{(n)}\) by (1.3) for each \(n \geq 1\) and replace \(y\) in (1.4) by \(\bar{y}^{(n)}\) to obtain
\[
\begin{align*}
x_t^{(n)} &= \arg \min_{x \in \mathcal{X}} \left\{ \mathcal{R}(x) - \langle \lambda_t^{(n)}, Ax \rangle \right\}, \\
\lambda_{t+1}^{(n)} &= \lambda_t^{(n)} - \gamma (Ax_t^{(n)} - \bar{y}^{(n)}) \\
&= (1.5)
\end{align*}
\]
with the initial guess \(\lambda_0^{(n)} = 0\) and a step size \(\gamma > 0\). This is the dual gradient method that we will consider for solving (1.2) using the average of multiple random measurement data. Since \(y_1, y_2, \ldots\) are random variables, we will consider the method (1.5) based on a stochastic analysis. We will establish convergence results by using the expectation of the Bregman distance induced by \(\mathcal{R}\) and hence the root mean square considering that \(\mathcal{R}\) is strongly convex.

When \(\mathcal{X}\) is a Hilbert space and \(\mathcal{R}(x) = \|x\|^2/2\), the method (1.5) becomes the Landweber iteration in Hilbert spaces using the average of multiple repeated measurement data which has been analyzed in [13] as a special case of a class of linear spectral regularization methods. The analysis in [13] is based on the spectral theory of bounded linear self-adjoint operators in Hilbert spaces. These tools however are no longer applicable to the method (1.5) due to
the possible non-Hilbertian structure of $\mathcal{X}$ and non-quadraticity of the regularization function $R$. In this paper we use tools of convex analysis in Banach spaces to analyze the method (1.5) with general Banach space $\mathcal{X}$ and general strongly convex function $R$. Without assuming any knowledge of the noise distribution, we obtain the convergence and rates of convergence for the method (1.5) as $n \to \infty$ when the method is terminated by either an a priori stopping rule or a modification of the discrepancy principle.

The paper is organized as follows, In section 2 we give a brief review of some basic facts from convex analysis in Banach spaces. In section 3 we provide the convergence analysis of the dual gradient method (1.5) when the method is terminated by either an a priori stopping rule or a modification of the discrepancy principle; in particular we derive the convergence rates when the sought solution satisfies certain variational source conditions. Finally in section 4, we provide various numerical results to test the performance of the proposed method.

2. Preliminaries

In this section, we collect some basic facts on convex analysis in Banach spaces which will be used in the analysis of the dual gradient method (1.5); for more information please refer to [42].

Let $\mathcal{X}$ be a Banach space, we use $\mathcal{X}^*$ to denote its dual space; in case $\mathcal{X}$ is a Hilbert space, $\mathcal{X}^*$ is identified with $\mathcal{X}$. Given $x \in \mathcal{X}$ and $\xi \in \mathcal{X}^*$ we write $\langle \xi, x \rangle = \xi(x)$ for the duality pairing. For a convex function $f : \mathcal{X} \to (-\infty, \infty]$, its effective domain is $\text{dom}(f) := \{x \in \mathcal{X} : f(x) < \infty\}$.

If $\text{dom}(f) \neq \emptyset, f$ is called proper. Given $x \in \text{dom}(f)$, an element $\xi \in \mathcal{X}^*$ is called a subgradient of $f$ at $x$ if

$$f(\bar{x}) \geq f(x) + \langle \xi, \bar{x} - x \rangle, \quad \forall \bar{x} \in \mathcal{X}.$$ 

The collection of all subgradients of $f$ at $x$ is denoted as $\partial f(x)$ and is called the subdifferential of $f$ at $x$. By definition we have

$$x^* \in \arg \min_{x \in \mathcal{X}} f(x) \iff 0 \in \partial f(x^*).$$  \hspace{1cm} (2.1)

Let

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

Given $x \in \text{dom}(\partial f)$ and $\xi \in \partial f(x)$, the Bregman distance induced by $f$ at $x$ in the direction $\xi$ is defined by

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

which is always nonnegative.

For a proper convex function $f : \mathcal{X} \to (-\infty, \infty]$, its convex conjugate is defined by

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

which is a convex function taking values in $(-\infty, \infty]$. By definition there holds the Fenchel–Young inequality

$$f^*(\xi) + f(x) \geq \langle \xi, x \rangle, \quad \forall x \in \mathcal{X}, \xi \in \mathcal{X}^*.$$  \hspace{1cm} (2.2)
If $f : \mathcal{X} \to (-\infty, \infty]$ is proper, lower semi-continuous and convex, so is $f^*$ and
\[
\xi \in \partial f(x) \iff x \in \partial f^*(\xi) \iff f(x) + f^*(\xi) = \langle \xi, x \rangle.
\tag{2.3}
\]

In applications, optimization problems of the form
\[
\inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\}
\]
are frequently encountered, where $A : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, and $f : \mathcal{X} \to (-\infty, \infty]$ and $g : \mathcal{Y} \to (-\infty, \infty]$ are proper convex functions. Let $A^* : \mathcal{Y}^* \to \mathcal{X}^*$ denote the adjoint of $A$. By the Fenchel–Young inequality we have
\[
f(x) \geq \langle \xi, x \rangle - f^*(\xi), \quad g(Ax) \geq (-\eta, Ax) - g^*(-\eta)
\]
and therefore
\[
f(x) + g(Ax) \geq \langle \xi - A^*\eta, x \rangle - f^*(\xi) - g^*(-\eta)
\]
for all $x \in \mathcal{X}$, $\xi \in \mathcal{X}^*$ and $\eta \in \mathcal{Y}^*$. Thus, taking $\xi = A^*\eta$ gives
\[
\inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\} \geq \sup_{\eta \in \mathcal{Y}^*} \{ -f^*(A^*\eta) - g^*(-\eta) \}.
\]

It is natural to ask under what conditions on $f$ and $g$ we have equality in the above equation. This is answered by the Fenchel–Rockafellar duality formula given below (see [42, corollary 2.8.5]).

**Proposition 2.1.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, let $f : \mathcal{X} \to (-\infty, \infty]$ and $g : \mathcal{Y} \to (-\infty, \infty]$ be proper, convex functions, and let $A : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. If there is $x_0 \in \text{dom}(f)$ such that $Ax_0 \in \text{dom}(g)$ and $g$ is continuous at $Ax_0$, then
\[
\inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\} = \sup_{\eta \in \mathcal{Y}^*} \{ -f^*(A^*\eta) - g^*(-\eta) \}.
\tag{2.4}
\]

A proper function $f : \mathcal{X} \to (-\infty, \infty]$ is called $c_0$-strongly convex for some constant $c_0 > 0$ if
\[
f(\bar{x}x + (1-s)x) + c_0 s (1-s) \|x - \bar{x}\|^2 \leq sf(\bar{x}) + (1-s)f(x)
\tag{2.5}
\]
for all $\bar{x}, x \in \text{dom}(f)$ and $s \in [0, 1]$. It is easy to show that for a proper $c_0$-strongly convex function $f : \mathcal{X} \to (-\infty, \infty]$ there holds
\[
D_f^2(\bar{x}, x) \geq c_0 \|x - \bar{x}\|^2
\tag{2.6}
\]
for all $\bar{x} \in \text{dom}(f)$, $x \in \text{dom}(\partial f)$ and $x \in \partial f(x)$. Furthermore, [42, corollary 3.5.11] contains the following important result concerning the differentiability of $f^*$ for strongly convex $f$.

**Proposition 2.2.** Let $\mathcal{X}$ be a Banach space and let $f : \mathcal{X} \to (-\infty, \infty]$ be a proper, lower semi-continuous, $c_0$-strongly convex function for some constant $c_0 > 0$. Then $\text{dom}(f^*) = \mathcal{X}^*$, $f^*$ is Fréchet differentiable and its gradient $\nabla f^*$ maps $\mathcal{X}^*$ into $\mathcal{X}$ with
\[
\|\nabla f^*(\xi) - \nabla f^*(\eta)\| \leq \frac{\|\xi - \eta\|}{2c_0}
\]
for all $\xi, \eta \in \mathcal{X}^*$. 

---
Given a proper strongly convex function $f: X \to (-\infty, \infty]$, we may consider for each $\xi \in X^*$ the convex minimization problem
\[
\min_{x \in X} \{ f(x) - \langle \xi, x \rangle \}
\] (2.7)
which is involved in the dual gradient method (1.5). According to [42, theorem 3.5.8], (2.1), (2.3) and proposition 2.2 we have

**Proposition 2.3.** If $f: X \to (-\infty, \infty]$ is a proper, lower semi-continuous, strongly convex function, then for any $\xi \in X^*$ the minimization problem (2.7) has a unique minimizer given by $\nabla f^*(\xi)$.

### 3. Convergence analysis

In this section we will provide convergence analysis of the method (1.5) when it is terminated by either an a priori stopping rule or a statistical variant of the discrepancy principle. We will make use of the following assumption.

**Assumption 3.1.**

(i) $X$ is a Banach space, $Y$ is a Hilbert space, and $A: X \to Y$ is a bounded linear operator.

(ii) $R: X \to [0, \infty]$ is proper, lower semi-continuous and $c_0$-strongly convex for some constant $c_0 > 0$.

(iii) The equation $Ax = y$ has a solution in $\text{dom}(R)$.

(iv) $y_1, y_2, \ldots$ is a sequence of independent identically distributed $Y$-valued random variables defined on a probability space $(\Omega, F, \mathbb{P})$ with $\mathbb{E}[y_1] = y$ and $0 < \sigma^2 := \mathbb{E}[\|y_1 - y\|^2] < \infty$.

The items (i)–(iii) in assumption 3.1 are standard. According to (i)–(iii) in assumption 3.1 and proposition 2.3, the equation (1.1) has a unique $R$-minimizing solution, denoted as $x^*$, and the method (1.5) is well-defined with
\[
A^* \lambda^{(n)} \in \partial \mathcal{R}(x^{(n)}) \quad \text{and} \quad x^{(n)}_t = \nabla \mathcal{R}^*(A^* \lambda^{(n)})
\] (3.1)
for each integer $t \geq 0$. For a sequence of noisy data $\{y_i\}$ satisfying (iv) in assumption 3.1, we define $\hat{y}^{(n)}$ by (1.3) for each $n \geq 1$. There holds
\[
\mathbb{E}\left[\left\|\hat{y}^{(n)} - y\right\|^2\right] = \frac{\sigma^2}{n}.
\] (3.2)

Our convergence analysis of the method (1.5), in particular the derivation of convergence rates, is based on the following result which has been established in [24].

**Proposition 3.1.** Let (i)–(iii) in assumption 3.1 be fulfilled and let $d_{\gamma^{(n)}}(\lambda) := \mathcal{R}^*(A^* \lambda) - \langle \lambda, \hat{y}^{(n)} \rangle$. Let $L := \|A\|^2/(2c_0)$. Consider the dual gradient method (1.5) with $\lambda_0^{(n)} = 0$. If $0 < \gamma < 1/L$ then for any $\lambda \in Y$ there holds
\[
d_{\gamma^{(n)}}(\lambda) - d_{\gamma^{(n)}}(\lambda_{t+1}^{(n)}) \geq \frac{1}{2\gamma(t+1)} \left( \| \lambda - \lambda_{t+1}^{(n)} \|^2 - \| \lambda \|^2 \right)
\]
\[
+ \left\{ \left( \frac{1}{2} - \frac{L\gamma}{4} \right) t + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \right\} \gamma \| A\lambda_{t}^{(n)} - \hat{y}^{(n)} \|^2
\]
for all integers $t \geq 0$. 


Based on proposition 3.1, we can obtain the following estimates which will be used frequently in the forthcoming analysis.

**Lemma 3.2.** Let (i)–(iii) in assumption 3.1 hold and consider the method (1.5) with $\lambda_0^{(a)} = 0$. Let $L := \|A\|^2/(2c_0)$ and assume $0 < \gamma < 1/L$. Then

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 + \frac{1}{8\gamma(t + 1)} \left\| \lambda^{(n)} t + 1 \right\|^2 \leq \eta_t + \gamma(t + 1) \left\| \tilde{y}^{(n)} - y \right\|^2$$

for all $t \geq 0$, where $c := 1/2 - L\gamma/2 > 0$ and

$$\eta_t := \sup_{x \in X} \left\{ \mathcal{R}(x) - \mathcal{R}(x) - \frac{1}{3} \gamma(t + 1)\|Ax - y\|^2 \right\}.$$

**Proof.** We use the similar argument in the proof of [24, theorem 3.4]. Let $d_\gamma(\lambda) := \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle$ and note that $d_\gamma(\lambda) = d_\gamma(\lambda) - \langle \lambda, \tilde{y}^{(n)} - y \rangle$. We may use proposition 3.1 to obtain for all $\lambda \in \mathcal{Y}$ that

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 \leq d_\gamma(\lambda) - d_\gamma(\lambda^{(n)} t + 1) + \left\langle \lambda^{(n)} t + 1 - \lambda, \tilde{y}^{(n)} - y \right\rangle \leq -\frac{1}{2\gamma(t + 1)} \left( \left\| \lambda^{(n)} t + 1 - \lambda \right\|^2 - \|\lambda\|^2 \right).$$

By using the Cauchy–Schwarz inequality to treat the term $\left\langle \lambda^{(n)} t + 1 - \lambda, \tilde{y}^{(n)} - y \right\rangle$ we further have

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 + \frac{1}{4\gamma(t + 1)} \left\| \lambda^{(n)} t + 1 - \lambda \right\|^2 \leq d_\gamma(\lambda) - d_\gamma(\lambda^{(n)} t + 1) + \frac{1}{4\gamma(t + 1)} \left\| \lambda \right\|^2 + \gamma(t + 1) \left\| \tilde{y}^{(n)} - y \right\|^2,$$

which together with the inequality $\left\| \lambda^{(n)} t + 1 \right\| \leq 2(\left\| \lambda^{(n)} t + 1 - \lambda \right\|^2 + \|\lambda\|^2)$ shows

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 + \frac{1}{8\gamma(t + 1)} \left\| \lambda^{(n)} t + 1 \right\|^2 \leq d_\gamma(\lambda) - d_\gamma(\lambda^{(n)} t + 1) + \frac{3\|\lambda\|^2}{4\gamma(t + 1)} + \gamma(t + 1) \left\| \tilde{y}^{(n)} - y \right\|^2. \quad (3.3)$$

By using the fact $y = Ax^t$ and the Fenchel–Young inequality (2.2), we have $d_\gamma(\lambda^{(n)} t + 1) \geq -\mathcal{R}(x^t)$. Therefore

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 + \frac{1}{8\gamma(t + 1)} \left\| \lambda^{(n)} t + 1 \right\|^2 \leq \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^t) + \frac{3\|\lambda\|^2}{4\gamma(t + 1)} + \gamma(t + 1) \left\| \tilde{y}^{(n)} - y \right\|^2$$

for all $\lambda \in \mathcal{Y}$ and hence

$$c\gamma(t + 1) \left\| A^{(n)} t - \tilde{y}^{(n)} \right\|^2 + \frac{1}{8\gamma(t + 1)} \left\| \lambda^{(n)} t + 1 \right\|^2 \leq \mathcal{R}(x^t) - \sup_{\lambda \in \mathcal{Y}} \left\{ -\mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle - \frac{3\|\lambda\|^2}{4\gamma(t + 1)} \right\} + \gamma(t + 1) \left\| \tilde{y}^{(n)} - y \right\|^2.$$
An application of the Fenchel–Rockafellar duality formula given in proposition 2.1 then completes the proof.

3.1. Convergence analysis under a priori stopping rule

In this subsection we consider the dual gradient method (1.5) and show that \( h_t^{(n)} \) converges to the unique \( R \)-minimizing solution \( x^* \) of (1.1) if \( t_n \) is chosen such that \( t_n \rightarrow \infty \) and \( t_n/n \rightarrow 0 \) as \( n \rightarrow \infty \). Furthermore we derive the convergence rates under suitable a priori choice of \( t_n \) when \( x^* \) satisfies certain variational source conditions.

By using [29, lemma 3.7] we immediately obtain the following convergence result for the sequence \( \{x_t, \lambda_t\} \) defined by the method (1.4) with exact data.

**Lemma 3.3.** Let (i)–(iii) in assumption 3.1 hold. Consider the method (1.4) with exact data. If \( \lambda_0 = 0 \) and \( 0 < \gamma < 4c_0/\|A\|^2 \), then \( D_{R_{\lambda}}^* \{x_t, \lambda_t\} \rightarrow 0 \) as \( t \rightarrow \infty \).

Although \( \{x_t, \lambda_t\} \) are deterministic, \( \{x_t^{(n)}, \lambda_t^{(n)}\} \) are random variables. Therefore, we need to consider \( \{x_t^{(n)}, \lambda_t^{(n)}\} \) by a stochastic analysis.

**Lemma 3.4.** Let assumption 3.1 hold. Let \( \{x_t, \lambda_t\} \) and \( \{x_t^{(n)}, \lambda_t^{(n)}\} \) be defined by (1.4) and (1.5) respectively with the initial guess \( \lambda_0 = \lambda_0^{(n)} = 0 \). Then for all integers \( t \geq 0 \) there hold

\[
E \left[ \|x_{t+1}^{(n)} - x_t\|^2 \right] \rightarrow 0 \quad \text{and} \quad E \left[ \|\lambda_{t+1}^{(n)} - \lambda_t\|^2 \right] \rightarrow 0
\]

as \( n \rightarrow \infty \).

**Proof.** We will use an induction argument. Since \( \lambda_0^{(n)} = \lambda_0 = 0 \), we have \( x_0^{(n)} = \nabla R^*(A^* \lambda_0^{(n)}) = \nabla R^*(A^* \lambda_0) = x_0 \). Thus the result is true for \( t = 0 \). Now we assume that the result is true for some \( t \geq 0 \). Then, by the definition of \( \lambda_{t+1}^{(n)} \) and \( \lambda_{t+1} \), we have

\[
\|\lambda_{t+1}^{(n)} - \lambda_{t+1}\|^2 = \|\lambda_{t+1}^{(n)} - \lambda_t - \gamma A (x_{t}^{(n)} - x_t) + \gamma (y^{(n)} - y)\|^2.
\]

By using the inequality \( (a + b + c)^2 \leq 3(a^2 + b^2 + c^2) \), taking the expectation, and using (3.2), we can obtain

\[
E \left[ \|\lambda_{t+1}^{(n)} - \lambda_{t+1}\|^2 \right] \\
\leq 3E \left[ \|\lambda_{t}^{(n)} - \lambda_t\|^2 \right] + 3\gamma^2 E \left[ \|A (x_t^{(n)} - x_t)\|^2 \right] + 3\gamma^2 E \left[ \|y^{(n)} - y\|^2 \right] \\
\leq 3E \left[ \|\lambda_t^{(n)} - \lambda_t\|^2 \right] + 3\gamma^2 E \left[ \|\lambda_t^{(n)} - x_t\|^2 \right] + \frac{3\gamma^2 \sigma^2}{n}.
\]

Thus, by the induction hypothesis we can conclude that \( E[\|\lambda_{t+1}^{(n)} - \lambda_{t+1}\|^2] \rightarrow 0 \) as \( n \rightarrow \infty \). 

Note that

\[
x_{t+1}^{(n)} = \nabla R^* \left( A^* \lambda_{t+1}^{(n)} \right) \quad \text{and} \quad x_{t+1} = \nabla R^* (A^* \lambda_{t+1}).
\]
By proposition 2.2 we have
\[
E \left[ \left\| x_{t+1} - x_t \right\|^2 \right] = E \left[ \left\| \nabla R^* (A^* \lambda_t^{(n)}) - \nabla R^* (A^* \lambda_{t+1}) \right\|^2 \right] \\
\leq \frac{1}{4c_0^2} E \left[ \left\| A^* (\lambda_t^{(n)}) - \lambda_{t+1} \right\|^2 \right] \\
\leq \frac{1}{4c_0} E \left[ \left\| \lambda_t^{(n)} - \lambda_{t+1} \right\|^2 \right] \\
\rightarrow 0
\]
as \( n \to \infty \). The proof is complete. \( \square \)

**Lemma 3.5.** Let assumption 3.1 hold. Consider the method (1.5) with \( \lambda_0^{(n)} = 0 \) and assume that \( 0 < \gamma < 4c_0/\|A\|^2 \). Let \( \Delta_t^{(n)} := D_{R^*}^{A^* \lambda_t^{(n)}} (x_t^{(n)}, y_t^{(n)}) \). Then
\[
E \left[ \Delta_t^{(n)} \right] - E \left[ \Delta_{t+1}^{(n)} \right] \leq C_1 \frac{\alpha^2}{n}
\]
for all integers \( t \geq 0 \), where \( C_1 := c_0 \gamma/(4c_0 - \gamma \|A\|^2) \).

**Proof.** Note that
\[
\Delta_t^{(n)} - \Delta_{t+1}^{(n)} = \left( \left\langle A^* \lambda_t^{(n)}, x_t^{(n)} - x_{t+1}^{(n)} \right\rangle - \nabla R^* (x_t^{(n)}) \right) + \left( \nabla R^* (x_{t+1}^{(n)}) - \left\langle A^* \lambda_{t+1}^{(n)}, x_{t+1}^{(n)} \right\rangle \right).
\]
By using the fact \( A^* \lambda_t^{(n)} \in \partial R(x_t^{(n)}) \) we have from (2.3) that
\[
\nabla R^* (x_t^{(n)}) + \nabla R^* (A^* \lambda_t^{(n)}) = \left\langle A^* \lambda_t^{(n)}, x_t^{(n)} \right\rangle
\]
for all \( t \geq 0 \). Therefore
\[
\Delta_t^{(n)} - \Delta_{t+1}^{(n)} = \left( \left\langle A^* \lambda_t^{(n)}, x_t^{(n)} - x_{t+1}^{(n)} \right\rangle - R^* (x_t^{(n)}) - R^* (A^* \lambda_t^{(n)}) \right) - \left( \left\langle A^* \lambda_{t+1}^{(n)}, x_{t+1}^{(n)} \right\rangle - R^* (A^* \lambda_{t+1}^{(n)}) \right).
\]
Since \( x_t^{(n)} = \nabla R^* (A^* \lambda_t^{(n)}) \), by using proposition 2.2 we can obtain
\[
\Delta_t^{(n)} - \Delta_{t+1}^{(n)} = \left( \left\langle A^* \lambda_t^{(n)}, x_t^{(n)} - x_{t+1}^{(n)} \right\rangle - \left\langle A^* \lambda_t^{(n)}, A^* \lambda_{t+1}^{(n)} \right\rangle - \nabla R^* \left( A^* \lambda_t^{(n)} \right) \right) + \left( \left\langle A^* \lambda_{t+1}^{(n)}, x_{t+1}^{(n)} \right\rangle - \left\langle A^* \lambda_t^{(n)}, x_t^{(n)} \right\rangle \right)
\]
\[
\leq \frac{1}{4c_0} \left\| A^* \lambda_t^{(n)} - A^* \lambda_{t+1}^{(n)} \right\|^2 + \left\langle A^* \lambda_t^{(n)}, x_{t+1}^{(n)} - x_t^{(n)} \right\rangle.
\]
According to the definition of \( \lambda_{t+1}^{(n)} \) it is easy to see that
\[
A^* \lambda_{t+1}^{(n)} - A^* \lambda_t^{(n)} = -\gamma A^* \left( Ax_t^{(n)} - y_t^{(n)} \right).
\]
Therefore
\[
\Delta_t^{(n)} - \Delta_{t+1}^{(n)} \leq \frac{\gamma^2}{4c_0} \left\| A^* \left( Ax_t^{(n)} - y_t^{(n)} \right) \right\|^2 - \gamma \left\langle Ax_t^{(n)} - y_t^{(n)}, Ax_t^{(n)} - y_t^{(n)} \right\rangle
\]
\[
\leq - \left( 1 - \frac{\gamma}{4c_0^2} \right) \gamma \left\| Ax_t^{(n)} - y_t^{(n)} \right\|^2 + \gamma \left\| y_t^{(n)} - y_t \right\| \left\| Ax_t^{(n)} - y_t^{(n)} \right\|.
\]
By using the Young’s inequality we have
\[
\gamma \|y^{(n)} - y\| \|Ax_t^{(n)} - y^{(n)}\| \leq \left(1 - \frac{\gamma \|A\|^2}{4c_0}\right) \gamma \|Ax_t^{(n)} - y^{(n)}\|^2 + C_1 \|y^{(n)} - y\|^2,
\]
with $C_1$ defined above. Therefore
\[
\Delta_{t+1}^{(n)} - \Delta_t^{(n)} \leq C_1 \|y^{(n)} - y\|^2.
\]
By taking the expectation and using (3.2) we then obtain
\[
E\left[\Delta_{t+1}^{(n)}\right] - E\left[\Delta_t^{(n)}\right] \leq C_1 E\left[\|y^{(n)} - y\|^2\right] = C_1 \frac{\sigma^2}{n}
\]
which shows the desired inequality. \hfill \Box

**Theorem 3.6.** Let assumption 3.1 hold. Consider the method (1.5) with $\lambda_0^{(n)} = 0$ and assume that $0 < \gamma < 4c_0 / \|A\|^2$. If $t_n$ is chosen such that $t_n \to \infty$ and $t_n/n \to 0$ as $n \to \infty$, then
\[
E\left[D_R^\lambda \left(x_t^{(n)} - x_n^{(n)}\right)\right] \to 0
\]
and hence $E\left[\|x_t^{(n)} - x_t\|^2\right] \to 0$ as $n \to \infty$.

**Proof.** By the strong convexity of $\mathcal{R}$ it suffices to show $E[\Delta_t^{(n)}] \to 0$ as $n \to \infty$. Let $t \geq 0$ be any fixed integer. Since $t_n \to \infty$ as $n \to \infty$, we have $t_n \geq t$ for large $n$. Thus, we may repeatedly use lemma 3.5 to obtain
\[
E\left[\Delta_t^{(n)}\right] \leq E\left[\Delta_t^{(n)}\right] + C_1(t_n - t) \frac{\sigma^2}{n}.
\]
Since $t_n/n \to 0$ as $n \to \infty$, we thus have
\[
\limsup_{n \to \infty} E\left[\Delta_t^{(n)}\right] \leq \limsup_{n \to \infty} E\left[\Delta_t^{(n)}\right] \quad (3.4)
\]
for any $t \geq 0$. Note that
\[
E\left[\Delta_t^{(n)}\right] = D_R^\lambda \left(x_t, x_t^{(n)}\right) = \left(\langle A^* \lambda_t, x_t - x_t^{(n)}\rangle - E\left[\langle A^* \lambda_t, x_t - x_t^{(n)}\rangle\right]\right) + \left(\mathcal{R}(x_t) - E\left[\mathcal{R}\left(x_t^{(n)}\right)\right]\right).
\]
With the help of the Cauchy–Schwarz inequality we have
\[
\begin{align*}
&\quad \left|E\left[\langle A^* \lambda_t^{(n)}, x_t - x_t^{(n)}\rangle\right] - \langle A^* \lambda_t, x_t - x_t^{(n)}\rangle\right| \\
&\leq \left|E\left[\langle A^* \lambda_t^{(n)} - \lambda_t, x_t - x_t^{(n)}\rangle\right]\right| + \left|E\left[\langle A^* \lambda_t, x_t - x_t^{(n)}\rangle\right]\right| \\
&\leq \|A\|\left(\left|E\left[\|\lambda_t^{(n)} - \lambda_t\|^2\right]\right|^{1/2}\left(\left|E\left[\|x_t - x_t^{(n)}\|^2\right]\right|\right)^{1/2}
\right.
\end{align*}
\]
\[
+ \|A^* \lambda_t\|\left(\left|E\left[\|x_t - x_t^{(n)}\|^2\right]\right|\right)^{1/2}. \quad (3.5)
\]
Thus we may use lemma 3.4 to conclude
\[
\lim_{n \to \infty} \mathbb{E} \left[\left( A^* \lambda^{(n)}_t, x^\dagger - x^{(n)}_t \right) \right] = \left( A^* \lambda^\dagger, x^\dagger - x_t \right). \tag{3.6}
\]
Next we will show that
\[
\mathcal{R}(x_t) \leq \liminf_{n \to \infty} \mathbb{E} \left[\mathcal{R} \left( x^{(n)}_t \right) \right]. \tag{3.7}
\]
To see this, we take a subsequence \( \{n_k\} \) with \( n_k \to \infty \) as \( k \to \infty \) such that
\[
\lim_{k \to \infty} \mathbb{E} \left[\mathcal{R} \left( x^{(n_k)}_t \right) \right] = \liminf_{n \to \infty} \mathbb{E} \left[\mathcal{R} \left( x^{(n)}_t \right) \right].
\]
According to lemma 3.4 we have \( \mathbb{E}[\|x^{(n_k)}_t - x_t\|^2] \to 0 \) as \( k \to \infty \). By taking a subsequence of \( \{n_k\} \) if necessary, we can guarantee \( \|x^{(n_k)}_t - x_t\| \to 0 \) as \( k \to \infty \) almost surely. Thus, from the lower semi-continuity of \( \mathcal{R} \) and Fatou’s lemma it follows
\[
\mathcal{R}(x_t) = \mathbb{E}[\mathcal{R}(x_t)] \leq \mathbb{E} \left[\liminf_{k \to \infty} \mathcal{R} \left( x^{(n_k)}_t \right) \right] = \liminf_{k \to \infty} \mathbb{E} \left[\mathcal{R} \left( x^{(n_k)}_t \right) \right]
\]
which shows (3.7). Consequently it follows from (3.5)–(3.7) that
\[
\limsup_{n \to \infty} \mathbb{E} \left[\Delta_t^{(n)} \right] \leq D^*_{\mathcal{R}} \left( x^\dagger, x_t \right).
\]
Therefore we may use (3.4) to obtain
\[
\limsup_{n \to \infty} \mathbb{E} \left[\Delta_t^{(n)} \right] \leq D^*_{\mathcal{R}} \left( x^\dagger, x_t \right)
\]
for all \( t \geq 0 \). Letting \( t \to \infty \) and using lemma 3.3 we thus obtain \( \mathbb{E}[\Delta_n^{(n)}] \to 0 \) as \( n \to \infty \). \( \square \)

We next consider deriving convergence rates of the method (1.5) under an \textit{a priori} stopping rule when the sought solution \( x^\dagger \) satisfies the variational source conditions specified in the following assumption.

\textbf{Assumption 3.2.} For the unique solution \( x^\dagger \) of (1.1) there is an error measure function \( \mathcal{E}^\dagger : \text{dom}(\mathcal{R}) \to [0, \infty) \) with \( \mathcal{E}^\dagger (x^\dagger) = 0 \) such that
\[
\mathcal{E}^\dagger (x) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + M\|Ax - y\|^q, \quad \forall x \in \text{dom}(\mathcal{R})
\]
for some \( 0 < q \leq 1 \) and some constant \( M > 0 \).

\textbf{Remark 3.1.} Variational source conditions were first introduced in [16], as a generalization of the spectral source conditions in Hilbert spaces, to derive convergence rates of Tikhonov regularization in Banach spaces. This kind of source conditions was further generalized and refined subsequently, see [9, 17, 18, 20] for instance. The error measure function \( \mathcal{E}^\dagger \) in assumption 3.2 is used to measure the speed of convergence; the usual choice of \( \mathcal{E}^\dagger \) is the Bregman distance induced by \( \mathcal{R} \).

\textbf{Remark 3.2.} When \( x^\dagger \) satisfies the benchmark source condition \( A^* \lambda^\dagger \in \partial \mathcal{R} (x^\dagger) \) for some \( \lambda^\dagger \in \mathcal{D} \), it is straightforward to see that
\[
D^*_{\mathcal{R}} \left( x^\dagger, x_t \right) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \|\lambda^\dagger\| \|Ax - y\|, \quad \forall x \in \text{dom}(\mathcal{R})
\]
which shows the variational source condition is satisfied with $E^\dagger(x) = D^{\dagger}_{\mathcal{R}}\lambda^\dagger(x, x^\dagger)$, $M = \|\lambda^\dagger\|$ and $q = 1$; this is a well-known fact, see [16].

**Remark 3.3.** When both $X$ and $Y$ are Hilbert spaces, $\mathcal{R}$ is $c_0$-strongly convex for some constant $c_0 > 0$, and $x^\dagger$ satisfies the source condition

$$\xi^\dagger := (A^*A)^{\nu/2} \omega \in \partial \mathcal{R} \left( x^\dagger \right)$$

for some $0 < \nu \leq 1$ and $\omega \in X$, then the variational source condition in assumption 3.2 holds with

$$E^\dagger(x) = \frac{1}{2} D^{\dagger}_{\mathcal{R}}(x, x^\dagger), \quad M = C_{\nu} \|\omega\|^{\frac{\nu}{1+\nu}} \quad \text{and} \quad q = \frac{2\nu}{1+\nu},$$

where $C_{\nu} := \frac{1+\nu}{2} \left( \nu \frac{1}{c_0} \right)^{\frac{1}{1+\nu}}$. Indeed, by the given condition (3.8), we can obtain

$$D^{\dagger}_{\mathcal{R}}(x, x^\dagger) = \mathcal{R}(x) - \mathcal{R}(x^\dagger) - \langle \omega, (A^*A)^{\nu/2} (x - x^\dagger) \rangle \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \|\omega\| \left\| (A^*A)^{\nu/2} (x - x^\dagger) \right\|.$$

By the interpolation inequality [8] we then have

$$D^{\dagger}_{\mathcal{R}}(x, x^\dagger) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \|\omega\| \|x - x^\dagger\|^{1-\nu} \|A (x - x^\dagger)\|^\nu.$$

Thus, an application of the Young’s inequality gives

$$D^{\dagger}_{\mathcal{R}}(x, x^\dagger) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \frac{1}{2} c_0 \|x - x^\dagger\|^2 + C_{\nu} \|\omega\|^{\frac{\nu}{1+\nu}} \|Ax - y\|^{\frac{\nu}{1+\nu}}.$$

Finally, by invoking the strong convexity of $\mathcal{R}$ and (2.6) we can obtain

$$D^{\dagger}_{\mathcal{R}}(x, x^\dagger) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \frac{1}{2} D^{\dagger}_{\mathcal{R}}(x, x^\dagger) + C_{\nu} \|\omega\|^{\frac{\nu}{1+\nu}} \|Ax - y\|^{\frac{\nu}{1+\nu}}$$

which implies the assertion.

It should be pointed out that the source condition (3.8) has been used in [11] to derive convergence rates of Tikhonov regularization in Hilbert spaces with non-quadratic penalty terms. We would also like to mention that for the special case $\mathcal{R}(x) = \frac{1}{2} \|x\|^2 + \nu\mathcal{E}(x)$, where $\nu\mathcal{E}$ denotes the indicator function of a closed convex set $\mathcal{E} \subset \mathcal{H}$, i.e. $\nu\mathcal{E}(x) = 0$ if $x \in \mathcal{E}$ and $\infty$ otherwise, the source condition (3.8) becomes the projected spectral source condition

$$x^\dagger = P_{\mathcal{E}}((A^*A)^{\nu/2} \omega)$$

for which it has been shown in [24] that the projected source condition implies the variational source condition. Here $P_{\mathcal{E}}$ denotes the metric projection of $\mathcal{H}$ onto $\mathcal{E}$.

**Remark 3.4.** The variational source conditions have been verified for various concrete inverse problems, see [6, 7, 18, 19] for instance.

**Theorem 3.7.** Let assumption 3.1 hold and consider the method (1.5) with $\lambda_0^{(n)} = 0$. Assume that $0 < \gamma < 2c_0/\|A\|^2$ and that $x^\dagger$ satisfies assumption 3.2. If $t_n$ is chosen such that $t_n \sim (n/\sigma^2)^{1+\nu}$, then
\[ E \left[ E^\dagger \left( x_n^{(n)} \right) \right] = O \left( \left( \frac{\sigma^2}{n} \right)^{q/2} \right). \tag{3.9} \]

Consequently, if there exists \( \lambda^i \in \mathcal{Y} \) such that \( A^* \lambda^i \in \partial \mathcal{R}(x^i) \), then with the choice \( t_n \sim \sqrt{n}/\sigma \) there hold

\[ E \left[ D^\dagger_R \lambda^i \left( x_n^{(n)} - x^i \right) \right] = O \left( \frac{\sigma}{\sqrt{n}} \right) \quad \text{and} \quad E \left[ \left\| x_n^{(n)} - x^i \right\|^2 \right] = O \left( \frac{\sigma}{\sqrt{n}} \right). \]

**Proof.** According to remark 3.2 and the strong convexity of \( \mathcal{R} \), it suffices to show (3.9). From lemma 3.2 and (3.2) it follows that

\[ c \gamma(t+1) E \left[ \left\| A x^{(n)} - y^{(n)} \right\|^2 \right] + \frac{1}{8 \gamma(t+1)} E \left[ \left\| \lambda^{(n)} \right\|^2 \right] \leq \eta_i + \gamma(t+1) E \left[ \left\| y^{(n)} - y \right\|^2 \right] = \eta_i + \gamma(t+1) \frac{\sigma^2}{n}. \]

By using the variational source condition and the nonnegativity of \( E^\dagger \), we have

\[ 0 \leq \mathcal{R}(x) - \mathcal{R}(x^i) + M \| Ax - y \|^q, \quad \forall x \in \mathcal{X}. \]

Therefore

\[ \eta_i \leq \sup_{x \in \mathcal{X}} \left\{ M \| Ax - y \|^q - \frac{1}{3} \gamma(t+1) \| Ax - y \|^2 \right\} \leq \sup_{\delta > 0} \left\{ M \delta^q - \frac{1}{3} \gamma(t+1) \delta^2 \right\} = c_2(t+1)^{-\frac{q}{2}} \sigma^2, \tag{3.10} \]

where \( c_2 := (1 - \frac{q}{2}) \left( \frac{3 M \delta}{2 \gamma} \right)^{\frac{q}{2}} M \). Consequently

\[ c \gamma(t+1) E \left[ \left\| A x^{(n)} - y^{(n)} \right\|^2 \right] + \frac{1}{8 \gamma(t+1)} E \left[ \left\| \lambda^{(n)} \right\|^2 \right] \leq c_2(t+1)^{-\frac{q}{2}} + \gamma(t+1) \frac{\sigma^2}{n}, \]

which shows that

\[ E \left[ \left\| A x^{(n)} - y^{(n)} \right\|^2 \right] \leq C \left( (t+1)^{-\frac{q}{2}} + \frac{\sigma^2}{n} \right), \]

\[ E \left[ \left\| \lambda^{(n)} \right\|^2 \right] \leq C \left( (t+1)^{\frac{q}{2}} + (t+1)^2 \frac{\sigma^2}{n} \right), \]

where here and below we use \( C \) to denote a generic constant independent of \( t, n, \) and \( \sigma \). Since \( t_n \) is chosen such that \( t_n \sim (n/\sigma^2)^{\frac{1}{2-q}} \), we have

\[ E \left[ \left\| A x^{(n)} - y^{(n)} \right\|^2 \right] \leq C \frac{\sigma^2}{n} \quad \text{and} \quad E \left[ \left\| \lambda^{(n)} \right\|^2 \right] \leq C \left( \frac{n}{\sigma^2} \right)^{1-q}. \]

The first estimate and (3.2) in particular imply

\[ E \left[ \left\| A x^{(n)} - y \right\|^2 \right] \leq 2E \left[ \left\| A x^{(n)} - y^{(n)} \right\|^2 \right] + 2E \left[ \left\| y^{(n)} - y \right\|^2 \right] \leq C \frac{\sigma^2}{n}. \]
Now we are ready to complete the proof. By using the variational source condition on $x^t$, the convexity of $\mathcal{R}$, and the fact $A^*\lambda_{t_n}^{(n)} \in \partial \mathcal{R} \left( x_{t_n}^{(n)} \right)$ we have

$$
\mathcal{E}^+ \left( x_{t_n}^{(n)} \right) \leq \mathcal{R} \left( x_{t_n}^{(n)} \right) - \mathcal{R} \left( x^t \right) + M \left\| A x_{t_n}^{(n)} - y \right\|^q
\leq \left\langle A^* \lambda_{t_n}^{(n)} , x_{t_n}^{(n)} - x^t \right\rangle + M \left\| A x_{t_n}^{(n)} - y \right\|^q
= \left\langle \lambda_{t_n}^{(n)} , A x_{t_n}^{(n)} - y \right\rangle + M \left\| A x_{t_n}^{(n)} - y \right\|^q
\leq \left\| \lambda_{t_n}^{(n)} \right\| \left\| A x_{t_n}^{(n)} - y \right\| + M \left\| A x_{t_n}^{(n)} - y \right\|^q.
$$

(3.11)

Therefore

$$
\mathbb{E} \left[ \mathcal{E}^+ \left( x_{t_n}^{(n)} \right) \right] \leq \mathbb{E} \left[ \left\| \lambda_{t_n}^{(n)} \right\| \left\| A x_{t_n}^{(n)} - y \right\| \right] + M \mathbb{E} \left[ \left\| A x_{t_n}^{(n)} - y \right\|^q \right]
\leq \left( \mathbb{E} \left[ \left\| \lambda_{t_n}^{(n)} \right\|^2 \right] \right)^{1/2} \left( \mathbb{E} \left[ \left\| A x_{t_n}^{(n)} - y \right\|^2 \right] \right)^{1/2}
+ M \left( \mathbb{E} \left[ \left\| A x_{t_n}^{(n)} - y \right\|^2 \right] \right)^{q/2}
\leq C \left( \frac{\sigma^2}{n} \right)^{q/2}.
$$

The proof is thus complete. \(\square\)

In theorem 3.7 we obtain a convergence rate result under an \textit{a priori} stopping rule. The rate depends on the variance $\sigma^2$, the number $n$ of samples, and the exponent $q$ in the variational source condition. For a fixed $q \in (0, 1]$, this rate illustrates the supplementary roles played by $\sigma^2$ and $n$: to achieve a prescribed error bound, if $\sigma$ is small then only a few number of samples are required; if $\sigma$ is large then a huge number of samples are required to suppress the variance in the averaged data. Note that in the variational source condition, the exponent $q$ is only allowed in $(0, 1]$ and thus the best possible rate we can expect is $O(\sigma / \sqrt{n})$. It is interesting to ask if it is possible to obtain a better convergence rate when the sought solution satisfies a higher order source condition. This question however remains open for general iterative regularization methods in Banach spaces. Finally it should be pointed out that the \textit{a priori} stopping rule used in theorem 3.7 requires the information of the variance $\sigma^2$ and the exponent $q$ in the source condition which are usually unknown. Therefore, the result obtained in theorem 3.7 has more theoretical interest than practical applications. In the next subsection we will consider an \textit{a posteriori} stopping rule for the method (1.5) without using any information on the exponent $q$ and the unknown variance $\sigma^2$.

### 3.2. Convergence analysis under a posteriori stopping rule

We next consider the dual gradient method (1.5) terminated by an \textit{a posteriori} stopping rule. The discrepancy principle is one of the most prominent rule that has been studied extensively. In this stopping rule, one needs the information on the noise level $\|y^{(n)} - y\|$. Since $y$ is unknown, we can not use this quantity directly. Recall that $\mathbb{E}[\|y^{(n)} - y\|^2] = \sigma^2/n$. To get an estimate on $\sigma^2$, we consider the square root of the sample variance.
\[ s_n := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}(n))}^2 \]

for which, according to (iv) in assumption 3.1, it is known that \( \mathbb{E}[s_n^2] = \sigma^2 \). Therefore, \( s_n^2 \) can be used as an estimator of the unknown variance \( \sigma^2 \) which is well known in statistics; see also [13] for a use in statistical inverse problems. Consequently, we may use \( s_n/\sqrt{n} \) as an estimator of \( \|\bar{y}(n) - y\| \). This leads us to propose the following stopping rule which is a statistical variant of the discrepancy principle.

**Rule 3.8.** Let \( \beta_0 > 0 \) be a given number. We define \( t_n \) to be the first integer such that
\[
\|Ax^{(n)}_t - y^{(n)}\| \leq \tau_n \frac{s_n}{\sqrt{n}},
\]

where \( \tau_n \geq 1 \) is a number that may depend on \( n \); if there is no such an integer \( t_n \) satisfying (3.12), we take \( t_n = \lfloor \beta_0n \rfloor \), where \( \lfloor \beta_0n \rfloor \) denotes the largest integer \( \leq \beta_0n \).

Recall that we have used \( s_n/\sqrt{n} \) as an estimator of \( \|\bar{y}(n) - y\| \). In case \( s_n/\sqrt{n} \) is an underestimated estimator, (3.12) may not be satisfied at a right number of iterations; continuing the iterations until (3.12) holds may result in a bad reconstruction result. Therefore, the requirement \( t_n \leq \beta_0n \) serves as an emergency stop. This kind of emergency stop has already been used in the analysis of finite dimensional approximations of regularization methods, see [25, 36] for instance; in the setting of statistical inverse problems it has been used in [2, 13, 27, 33]. It should be emphasized that \( t_n \) produced by rule 3.8 is a random variable instead of a deterministic quantity; this is a key difference between rule 3.8 and the classical discrepancy principle.

In order to carry out the analysis of the method (1.5) under rule 3.8, the key part is to construct an event \( \Omega_n \) with \( \mathbb{P}(\Omega_n) \to 1 \) as \( n \to \infty \) such that it can interplay well with rule 3.8. For linear regularization methods in Hilbert spaces, such an event has been constructed in [13] by using the data and the filter functions of the methods and, as a consequence, \( \tau_n \) in rule 3.8 is allowed to take as the constant 1; the linearity of the methods plays a crucial role for such a construction. In our situation, the regularized solution does not depend on the data linearly and the construction in [13] is no longer available. Instead, we will use only the noisy data to construct such an event \( \Omega_n \); in order for it to interplay with rule 3.8 well, we have to sacrifice a little bit to require \( \tau_n \to \infty \) slowly as \( n \to \infty \). In the forthcoming convergence analysis, we require the number \( \tau_n \) in rule 3.8 to satisfy
\[
\tau_n \to \infty \quad \text{and} \quad \tau_n/\sqrt{n} \to 0 \quad \text{as} \quad n \to \infty.
\]

For instance, we may take
\[
\tau_n := \max \{ \tau_0, \log |\log |\log n|| \},
\]

where \( \tau_0 \geq 1 \) is a fixed number. Based on \( \tau_n \), we define the event
\[
\Omega_n := \left\{ |s_n - \sigma| \leq \frac{\sigma}{2}, \|y^{(n)} - y\| \leq \tau_n \sqrt{\frac{\epsilon^2}{2n}} \right\}.
\]
on which we will perform the convergence analysis, where \( c > 0 \) is the constant appearing in lemma 3.2. We claim that

\[
P(\Omega_n) \to 1 \quad \text{as} \quad n \to \infty.
\]  

(3.15)

To see this, we note that

\[
s_n^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} \|y_i\|^2 - \left\| \bar{y}^{(n)} \right\|^2 \right).$

Since \( y_1, y_2, \ldots \) is a sequence of independent identically distributed random variables as assumed in (iv) of assumption 3.1, by the strong law of large numbers [32, corollary 7.10] we have

\[
\left\| \bar{y}^{(n)} \right\|^2 \to \|y\|^2 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \|y_i\|^2 \to E[\|y\|^2]
\]

as \( n \to \infty \) almost surely. Therefore

\[
s_n^2 \to E[\|y\|^2] - \|y\|^2 = \sigma^2
\]

and hence \( s_n \to \sigma \) as \( n \to \infty \) almost surely. Since almost sure convergence implies convergence in probability, we have

\[
P\left( |s_n - \sigma| \leq \frac{\sigma}{2} \right) \to 1 \quad \text{as} \quad n \to \infty.
\]  

(3.16)

By the Chebyshev’s inequality, (3.2), and the assumption \( \tau_n \to \infty \), we also have

\[
P\left( |s_n - \sigma| \leq \frac{\sigma}{2}, \left\| \bar{y}^{(n)} - y \right\| > \tau_n \sqrt{\frac{c \sigma^2}{2n}} \right) \leq P\left( \left\| \bar{y}^{(n)} - y \right\| > \sigma \tau_n \sqrt{\frac{c}{8n}} \right) \leq \frac{8n}{c \sigma^2 \tau_n^2} E\left[ \left\| \bar{y}^{(n)} - y \right\|^2 \right] \to 0
\]

as \( n \to \infty \). Therefore, by using (3.16), we have

\[
P(\Omega_n) = P\left( |s_n - \sigma| \leq \frac{\sigma}{2} \right) - P\left( |s_n - \sigma| \leq \frac{\sigma}{2}, \left\| \bar{y}^{(n)} - y \right\| > \tau_n \sqrt{\frac{c \sigma^2}{2n}} \right) \to 1
\]

as \( n \to \infty \).

The following result gives an upper bound estimate of \( t_n \) on \( \Omega_n \) defined by rule 3.8.

**Lemma 3.9.** Let assumption 3.1 hold and consider the method (1.5) with \( \lambda_0^{(n)} = 0 \). Assume that \( 0 < \gamma < 2c \sigma / \|A\|^2 \). Let \( t_n \) be defined by rule 3.8 with \( \tau_n \) satisfying (3.13) and set \( \phi(n) := \frac{\sigma \tau_n}{\sqrt{n}} \).

Then for any \( \varepsilon > 0 \) there is an integer \( n_c \) such that

\[
t_n \phi(n)^2 \leq \varepsilon \quad \text{on} \quad \Omega_n
\]

for all \( n \geq n_c \).

**Proof.** Note that on \( \Omega_n \) we always have \( \sigma / 2 \leq s_n \leq 3 \sigma / 2 \). If \( t_n \leq 1 / \phi(n) \), then \( t_n \phi(n)^2 \leq \phi(n) \).

In the following we will assume \( t_n > 1 / \phi(n) \). By the definition of \( t_n \) we have

\[
\|Ax_n^{(n)} - \bar{y}^{(n)}\| > \tau_n \frac{s_n}{\sqrt{n}}.
\]
Therefore, by using (3.3) we have on $\Omega_n$ that

$$
\begin{align*}
&\frac{e^{\gamma t_n} s_n^2}{n} + \frac{1}{8\gamma t_n} \left\| \lambda_a^{(n)} \right\|^2 \leq e^{\gamma t_n} \left\| A\lambda_a^{(n-1)} - y^{(n)} \right\|^2 + \frac{1}{8\gamma t_n} \left\| \lambda_a^{(n)} \right\|^2 \\
&\quad \leq d_s(\lambda) - d_s(\lambda_a^{(n)}) + \frac{3\left\| \lambda \right\|^2}{4\gamma t_n} + \gamma t_n \left\| y^{(n)} - y \right\|^2 \\
&\quad \leq d_s(\lambda) - d_s(\lambda_a^{(n)}) + \frac{3\left\| \lambda \right\|^2}{4\gamma t_n} + c\gamma t_n \frac{e^{\gamma t_n} s_n^2}{2n}
\end{align*}
$$

which implies that

$$
\begin{align*}
&\frac{c}{2} e^{\gamma t_n} \phi(n)^2 + \frac{1}{8\gamma t_n} \left\| \lambda_a^{(n)} \right\|^2 \leq \frac{c}{2} e^{\gamma t_n} \frac{e^{\gamma t_n} s_n^2}{2n} + \frac{1}{8\gamma t_n} \left\| \lambda_a^{(n)} \right\|^2 \\
&\quad \leq d_s(\lambda) - d_s(\lambda_a^{(n)}) + \frac{3\left\| \lambda \right\|^2}{4\gamma t_n}
\end{align*}
$$

(3.17)

for all $\lambda \in \mathcal{Y}$. Note that

$$
d_s(\lambda) = \mathcal{R}^*(A^*\lambda - \langle A^*\lambda, x^* \rangle) \geq -\mathcal{R}(x^*)
$$

which shows that

$$
\inf_{\lambda \in \mathcal{Y}} d_s(\lambda) \geq -\mathcal{R}(x^*) > -\infty.
$$

Therefore, for any $\varepsilon > 0$ we can find $\lambda_\varepsilon \in \mathcal{Y}$ such that

$$
d_s(\lambda_\varepsilon) \leq \inf_{\lambda \in \mathcal{Y}} d_s(\lambda) + \frac{c\gamma}{4} \varepsilon \leq d_s(\lambda_a^{(n)}) + \frac{c\gamma}{4} \varepsilon.
$$

This and (3.17) with $\lambda = \lambda_\varepsilon$ show that

$$
\frac{c}{2} e^{\gamma t_n} \phi(n)^2 + \frac{1}{8\gamma t_n} \left\| \lambda_a^{(n)} \right\|^2 \leq \frac{c}{2} e^{\gamma t_n} \varepsilon + \frac{3\left\| \lambda \varepsilon \right\|^2}{4\gamma t_n}
$$

(3.18)

which together with $t_n > 1/\phi(n)$ in particular implies

$$
t_n \phi(n)^2 \leq \frac{\varepsilon}{2} + \frac{3\left\| \lambda \varepsilon \right\|^2}{2c\gamma t_n} \leq \frac{\varepsilon}{2} + \frac{3\left\| \lambda \varepsilon \right\|^2}{2c\gamma^2} \phi(n).
$$

Therefore, we always have

$$
t_n \phi(n)^2 \leq \max \left\{ \phi(n), \frac{\varepsilon}{2} + \frac{3\left\| \lambda \varepsilon \right\|^2}{2c\gamma^2} \phi(n) \right\} \text{ on } \Omega_n.
$$

Since $\phi(n) \to 0$ as $n \to \infty$, we can find $n_\varepsilon$ such that $\phi(n) \leq \varepsilon$ and $\frac{3\left\| \lambda \varepsilon \right\|^2}{2c\gamma^2} \phi(n) \leq \frac{\varepsilon}{2}$ for all $n \geq n_\varepsilon$. Thus $t_n \phi(n)^2 \leq \varepsilon$ on $\Omega_n$ for all $n \geq n_\varepsilon$. \qed
Based on lemma 3.9, we are now ready to show a convergence result for $x_{k}^{(n)}$ with $t_n$ determined by rule 3.8.

**Theorem 3.10.** Let assumption 3.1 hold and consider the method (1.5) with $\lambda_0^{(n)} = 0$. Assume that $0 < \gamma < 2c_0 / \|A\|^2$. Let $t_n$ be determined by rule 3.8 with $\tau_n$ satisfying (3.13). Assume that there exists $\xi^\dagger \in \partial \mathcal{R}(x^\dagger)$ such that $\xi^\dagger \in \mathcal{A}^*(\mathcal{Y})$. Then for any $0 < \alpha < 1$ there holds

$$
E \left[ \left( D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) \right)^{\alpha} \right] \to 0 \quad \text{as} \quad n \to \infty
$$

(3.19)

and consequently for any $0 < p < 2$ there holds

$$
E \left[ \left\| x_{k}^{(n)} - x^\dagger \right\|^p \right] \to 0 \quad \text{as} \quad n \to \infty.
$$

(3.20)

**Proof.** First we have

$$
D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) \leq D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) + D_{R}^{A^*\lambda_{n}^{(n)}} (x_{k}^{(n)}, x^\dagger)
$$

$$
= \left\langle A^*\lambda_{n}^{(n)} - \xi^\dagger, x_{k}^{(n)} - x^\dagger \right\rangle.
$$

(3.21)

Since $\xi^\dagger \in \mathcal{A}^*(\mathcal{Y})$, for any $\varepsilon > 0$ we can find $\lambda_{n}^\dagger \in \mathcal{Y}$ and $\nu_{c} \in \mathcal{R}^*$ such that

$$
\xi^\dagger = A^*\lambda_{n}^\dagger + \nu_{c} \quad \text{and} \quad \|\nu_{c}\|^2 \leq \frac{1}{4} c_{0} \varepsilon.
$$

Therefore

$$
D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) \leq \left\langle A^*\lambda_{n}^{(n)} - A^*\lambda^\dagger - \nu_{c}, x_{k}^{(n)} - x^\dagger \right\rangle
$$

$$
\leq \left\langle \lambda_{n}^{(n)} - \lambda^\dagger, Ax_{k}^{(n)} - y \right\rangle + \|\nu_{c}\| \left\| x_{k}^{(n)} - x^\dagger \right\|.
$$

By using the strong convexity of $\mathcal{R}$ and the estimate on $\|\nu_{c}\|$ we have

$$
\|\nu_{c}\| \left\| x_{k}^{(n)} - x^\dagger \right\| \leq \frac{1}{2} c_{0} \left\| x_{k}^{(n)} - x^\dagger \right\|^2 + \frac{\|\nu_{c}\|^2}{2 c_{0}} \leq \frac{1}{2} D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) + \frac{\varepsilon}{8}.
$$

Consequently

$$
D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) \leq 2 \left\langle \lambda_{n}^{(n)} - \lambda^\dagger, Ax_{k}^{(n)} - y \right\rangle + \frac{\varepsilon}{4}
$$

$$
\leq 2 \left( \left\| \lambda_{n}^{(n)} \right\| + \left\| \lambda^\dagger \right\| \right) \left\| Ax_{k}^{(n)} - y \right\| + \frac{\varepsilon}{4}.
$$

(3.22)

We now show that there exists $n_{c}$ such that

$$
D_{R}^{A^*\lambda_{n}^{(n)}} (x^\dagger, x_{k}^{(n)}) \leq \varepsilon \quad \text{on} \quad \Omega_{n}
$$

(3.23)

for all $n \geq n_{c}$. According to lemma 3.9 and (3.13) we have $t_n < 1/\phi(n)^2 \leq [\beta_{0} n]$ on $\Omega_{n}$ for sufficiently large $n$. Therefore

$$
\left\| Ax_{k}^{(n)} - y^{(n)} \right\| \leq \tau_{n} \frac{S_{n}}{\sqrt{n}}
$$

Recently we have realized that this extra condition can be removed.
and consequently
\[
\|Ax_t^{(n)} - y\| \leq |Ax_t^{(n)} - y^{(n)}| + \|y^{(n)} - y\| \leq \tau_0 \frac{s_t}{\sqrt{n}} + \tau_0 \sqrt{\frac{c^2}{2n}} \\
\leq \frac{3\sigma\tau_0}{2\sqrt{n}} + \frac{3\sigma\tau_0}{2} \sqrt{\frac{c}{2n}} = 3 \left(1 + \sqrt{\frac{c}{2}}\right) \phi(n) \\
< \frac{9}{2} \phi(n).
\] (3.24)

By using (3.18) we have
\[
\|\lambda_t^{(n)}\|^2 \leq 2c\gamma^2 \epsilon_t + 6 \|\lambda_t\|^2,
\]
where \(\lambda_t \in \mathcal{Y}\) is an element chosen in the proof of lemma 3.9. Therefore, it follows from (3.22) that
\[
D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)}) \leq 9 \left(\sqrt{2c\gamma^2 \epsilon_t} + 6 \|\lambda_t\|^2 + \|\lambda_t\|^2\right) \phi(n) + \frac{\epsilon}{4}
\] (3.25)
on \(\Omega_n\). By using lemma 3.9 and the property \(\phi(n) \to 0\) as \(n \to \infty\), we may find a sufficiently large \(n_\epsilon\) such that
\[
\left(\sqrt{2c\gamma^2 \epsilon_t} + 6 \|\lambda_t\|^2 + \|\lambda_t\|^2\right) \phi(n) \leq \frac{\epsilon}{12}
\]
for all \(n \geq n_\epsilon\). Combining this with (3.25) we thus obtain (3.23).

Next we will show (3.19). Let \(\chi_{\Omega_n}\) denote the characteristic function of \(\Omega_n\), i.e. \(\chi_{\Omega_n}(\omega) = 1\) if \(\omega \in \Omega_n\) and \(\chi_{\Omega_n}(\omega) = 0\) otherwise. By using (3.23) and the Hölder inequality we have for sufficiently large \(n\) that
\[
E \left[\left(D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)})\right)\right]^\alpha \\
= E \left[D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)}) \chi_{\Omega_n}\right] + E \left[D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)}) \chi_{\Omega_n}\right]^\alpha \\
\leq \epsilon^\alpha + \left(E \left[D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)})\right]^\alpha\right) \mathbb{P}(\Omega_n^c)^{1-\alpha}.
\] (3.26)

If we are able to show
\[
E \left[D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)})\right] \leq C
\] (3.27)
for some constant \(C\) independent of \(n\), then, by using the fact that \(\mathbb{P}(\Omega_n^c) \to 0\) as \(n \to \infty\), we can conclude
\[
\limsup_{n \to \infty} E \left[D_{\mathcal{R}}^{A_{x_t}^{(n)}}(\chi_t^{(n)}, \chi_t^{(n)})\right]^{\alpha} \leq \epsilon^\alpha
\]
and thus obtain (3.19), due to the arbitrariness of \(\epsilon\).

It remains only to show (3.27). From (3.21), the Cauchy–Schwarz inequality and the strong convexity of \(\mathcal{R}\) it follows that
\[
D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)}) \leq \left\langle \lambda_{h}^{(n)},Ax_{h}^{(n)} - y \right\rangle + \|\xi^{\dagger}\|\|\lambda_{h}^{(n)} - x^{\dagger}\|
\leq \|\lambda_{h}^{(n)}\|\|Ax_{h}^{(n)} - y\| + \frac{1}{2c_{0}}\|\xi^{\dagger}\|^{2} + \frac{1}{2}D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)})
\]
which implies
\[
D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)}) \leq 2\|\lambda_{h}^{(n)}\|\|Ax_{h}^{(n)} - y\| + \frac{1}{c_{0}}\|\xi^{\dagger}\|^{2}.
\]

We next use lemma 3.2. By the definition of \(\eta_{t}\) and the nonnegativity of \(\mathcal{R}\), we have \(0 \leq \eta_{t} \leq \mathcal{R}(x^{\dagger}) < \infty\) for all \(t \geq 0\). Thus, it follows from lemma 3.2 that
\[
\|Ax_{h}^{(n)} - y^{(n)}\| \leq C\left(\frac{1}{t_{n}} + \|y^{(n)} - y\|\right), \quad \|\lambda_{h}^{(n)}\| \leq C\left(t_{n} + t_{n}^{2}\|\lambda^{(n)} - y\|\right).
\]

Therefore, by using \(t_{n} \leq \beta_{0}n\), we have
\[
D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)}) \leq C\left(\sqrt{t_{n}} + t_{n}\|y^{(n)} - y\|\right)\left(\frac{1}{\sqrt{t_{n}}} + \|y^{(n)} - y\|\right) + \frac{1}{c_{1}}\|\xi^{\dagger}\|^{2}
\leq C\left(1 + \sqrt{n}\|y^{(n)} - y\| + t_{n}\|y^{(n)} - y\|\right) + \frac{1}{c_{1}}\|\xi^{\dagger}\|^{2}
\leq C\left(1 + \sqrt{n}\|y^{(n)} - y\| + n\|y^{(n)} - y\|\right) + \frac{1}{c_{1}}\|\xi^{\dagger}\|^{2}.
\]
(3.28)

By taking the expectation and using (3.2), we can obtain
\[
\mathbb{E}\left[D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)})\right]
\leq C\left(1 + \sqrt{n}\left(\mathbb{E}\left[\|y^{(n)} - y\|^{2}\right]\right)^{1/2} + n\mathbb{E}\left[\|y^{(n)} - y\|^{2}\right]\right) + \frac{1}{c_{1}}\|\xi^{\dagger}\|^{2}
\leq C\left(1 + \sigma + \sigma^{2}\right) + \frac{1}{c_{1}}\|\xi^{\dagger}\|^{2}
\]
which shows (3.27). The proof is therefore complete. 

\[\square\]

**Remark 3.5.** In theorem 3.10 we have obtained the convergence result (3.19) with \(0 < \alpha < 1\). If in addition
\[
\sigma_{4} := \mathbb{E}[\|y_{1} - y\|^{4}] < \infty,
\]
we can improve the convergence result (3.19) to include \(\alpha = 1\) and hence \(\mathbb{E}\left[\|x_{h}^{(n)} - x^{\dagger}\|^{2}\right] \to 0\) as \(n \to \infty\). To see this, we may use the similar argument for deriving (3.26) to obtain
\[
\mathbb{E}\left[D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)})\right] \leq \varepsilon + \left(\mathbb{E}\left[\left(D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)})\right)^{2}\right]\right)^{1/2} \mathbb{P}(\Omega_{n})^{1/2}.
\]
(3.29)

Therefore, it suffices to show
\[
\mathbb{E}\left[\left(D_{\mathcal{R}}^{\ast\lambda^{(n)}}(x^{\dagger},x_{h}^{(n)})\right)^{2}\right] \leq C
\]
for some constant $C$ independent of $n$. By virtue of (3.28) it is easy to derive
\[
E \left[ \left( D^{\ast}_R \lambda^{(n)} (x^t, x^{(n)}) \right)^2 \right] 
\leq C \left( 1 + nE \left[ \|\hat{y}^{(n)} - y\|^2 \right] + n^2 E \left[ \|\hat{y}^{(n)} - y\|^4 \right] \right) + \frac{2}{c_1} \|\varepsilon^\dagger\|^4.
\]  
(3.30)

With the similar argument in the proof of [13, corollary 3], we can derive
\[
E \left[ \|\hat{y}^{(n)} - y\|^4 \right] \leq \frac{1}{n^3} E \left[ \|y_1 - y\|^4 \right] + \frac{3(n-1)}{n^3} \left( E \left[ \|y_1 - y\|^2 \right] \right)^2 
\leq \frac{\sigma t}{n^3} + \frac{3(n-1)}{n^3} \sigma^4.
\]

By invoking this estimate and (3.2), we can obtain (3.29) from (3.30) immediately.

**Theorem 3.11.** Let assumption 3.1 hold and consider the method (1.5) with $\lambda_0^{(n)} = 0$. Assume that $0 < \gamma < 2c^2/|A|^2$. Let $t_n$ be determined by rule 3.8 with $\tau_n$ satisfying (3.13). If $x^\dagger$ satisfies the variational source condition given in assumption 3.2, then there is a constant $C$ such that
\[
P \left( E^\dagger (x^{(n)}_t) \leq C \phi(n)^q \right) \to 1
\]
as $n \to \infty$, where $\phi(n) := \frac{\sqrt{s_2}}{2 \sqrt{n}}$ is defined in lemma 3.9.

**Proof.** Since $P(\Omega_n) \to 1$ as $n \to \infty$, it suffices to establish
\[
E^\dagger (x^{(n)}_t) \leq C \phi(n)^q \quad \text{on } \Omega_n
\]
for some constant $C$ independent of $n$. Under the given variational source condition on $x^\dagger$, we have the estimate (3.10) on $\Omega_n$. Combining this with lemma 3.2 shows that
\[
c_1 \gamma \left\| Ax^{(n)} - \hat{y}^{(n)} \right\| \leq c_2 t_n \sqrt{\frac{\gamma}{n}} + c_1 \gamma \left\| \hat{y}^{(n)} - y \right\| \leq c_2 t_n \sqrt{\frac{\gamma}{n}} + c_1 \gamma \frac{\tau_n^s s_2}{2n}
\]  
(3.31)

and
\[
\left\| \lambda^{(n)}_t \right\|^2 \leq 8 \gamma t_n \frac{s_2}{2n} + 8 \gamma t_n \frac{s_2}{2n} \left\| \hat{y}^{(n)} - y \right\|^2 \leq 8 \gamma t_n \frac{s_2}{2n} + 4c_2 \gamma \frac{\tau_n^s s_2}{2n}
\]  
(3.32)
on $\Omega_n$. Since $\left\| Ax^{(n)}_t - \hat{y}^{(n)} \right\| > \tau_n \sqrt{\frac{s_n}{n}}$, we have from (3.31) that
\[
c_1 \gamma \frac{s_2}{n} \leq c_2 t_n \sqrt{\frac{\gamma}{n}} + c_1 \gamma \frac{s_2}{2n}
\]
which together with $s_n \geq \sigma/2$ on $\Omega_n$ implies
\[
\frac{1}{2} c_1 \gamma \phi(n)^2 \leq c_2 t_n \sqrt{\frac{\gamma}{n}}.
\]
Therefore
\[
t_n \leq c_3 \phi(n)^{-2}
\]  
(3.33)
where $c_3 > 0$ is a constant independent of $n$. This in particular shows that $t_n < \beta_0 n$ and hence
\[
\left\| Ax^{(n)}_t - \hat{y}^{(n)} \right\| \leq \frac{\tau_n}{\sqrt{n}} \sqrt{n} \quad \text{on } \Omega_n \quad \text{for sufficiently large } n.
\]
Consequently
\[
\left\| Ax^{(n)}_t - y \right\| \leq \frac{9}{2} \phi(n)
\]
as argued in (3.24). By using (3.32), (3.33) and $s_n \leq 3\sigma/2$ we also have
\[
\left\| \lambda_t^{(n)} \right\|^2 \leq 8\gamma t_n^{3(1-q)} + 36\epsilon_1^2 n^2 \varphi(n)^2 \leq c_4 \phi(n)^2(1-q)
\]
for some constant $c_4$ independent of $n$. Finally we may use (3.11) to obtain
\[
\mathcal{E}_t^\dagger \left( x_t^{(n)} \right) \leq \left\| \lambda_t^{(n)} \right\| \|Ax_t^{(n)} - y\| + C\|Ax_t^{(n)} - y\|^q \leq c_5 \phi(n)^q
\]
for some constant $c_5$ independent of $n$. The proof is complete.

**Remark 3.6.** According to the definition of $\phi(n)$, theorem 3.11 gives the convergence rate
\[
\mathbb{P}\left( \mathcal{E}_t^\dagger \left( x_t^{(n)} \right) \leq C\frac{\sigma^q}{n} \left( \frac{\sigma^2}{n} \right)^{q/2} \right) \to 1 \quad \text{as } n \to \infty.
\]
Because $\tau_n \to \infty$, this rate is worse than the one derived in theorem 3.7 under an *a priori* stopping rule. The requirement $\tau_n \to \infty$ is used to construct $\Omega_n$ with $\mathbb{P}(\Omega_n) \to 1$ as $n \to \infty$. If such an even $\Omega_n$ could be constructed without using the requirement on $\tau_n$, the convergence rate in theorem 3.11 could be upgraded to $\mathbb{P}(\mathcal{E}_t^\dagger(x_t^{(n)}) \leq C(\sigma^2/n)^{q/2}) \to 1$ as $n \to \infty$. This question however remains open.

### 4. Numerical results

In this section we will report various numerical results to test the performance of the method (1.5).

**Example 4.1.** We first consider the application of the method (1.5) to solve linear ill-posed problems in Hilbert spaces with convex constraint. Let $A : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator between two Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and let $\mathcal{C} \subset \mathcal{X}$ be a closed convex set. Given $y \in A(\mathcal{C})$, we consider finding the unique solution $x^*$ of $Ax = y$ in $\mathcal{C}$ with minimal norm which can be stated as (1.2) with
\[
\mathcal{R}(x) := \frac{1}{2} \|x\|^2 + \nu_{\mathcal{C}}(x),
\]
where $\nu_{\mathcal{C}}$ denotes the indicator function of $\mathcal{C}$. Clearly $\mathcal{R}$ satisfies assumption 3.1(ii) with $c_0 = 1/2$. It is easy to see that the method (1.5) takes the form
\[
x_t^{(n)} = P_{\mathcal{C}}(A^* \lambda_t^{(n)}), \quad \lambda_{t+1}^{(n)} = \lambda_t^{(n)} - \gamma(Ax_t^{(n)} - y(n)),
\]
where $P_{\mathcal{C}}$ denotes the metric projection of $\mathcal{X}$ onto $\mathcal{C}$. In case $\mathcal{X} = L^2(\mathcal{D})$ for some domain $\mathcal{D} \subset \mathbb{R}^d$ and $\mathcal{C} = \{x \in \mathcal{X} : x \geq 0 \text{ a.e. on } \mathcal{D}\}$, the iteration scheme (4.1) becomes
\[
x_t^{(n)} = \max \left\{ A^* \lambda_t^{(n)}, 0 \right\}, \quad \lambda_{t+1}^{(n)} = \lambda_t^{(n)} - \gamma \left( Ax_t^{(n)} - y(n) \right),
\]
with initial guess $\lambda_0^{(n)} = 0$. This method can be viewed as an application of the Landweber iteration to the dual variable with nonnegative constraint on the primal variable.

We now test the performance of the method (4.2) by reconstructing nonnegative solutions of linear ill-posed problems and compare it with the Landweber iteration
\[
x_{t+1}^{(n)} = x_t^{(n)} - \gamma A^* \left( Ax_t^{(n)} - y(n) \right)
\]
that is considered in [13] which does not incorporate the nonnegative constraint. Let us consider the first kind Fredholm integral equation of the form

$$\int_0^1 k(s, s') x(s') \, dt = y(s) \quad \text{on } [0, 1]$$

with the kernel

$$k(s, s') = \begin{cases} 
40s(1-s'), & s \leq s' \\
40s'(1-s), & s \geq s'. 
\end{cases}$$

We assume the sought solution is given by

$$x^1(s) = \begin{cases} 
20s(s-0.2)(0.7-s), & 0.2 \leq s \leq 0.7, \\
0, & \text{elsewhere}
\end{cases}$$

which is nonnegative.

In practical applications, the data are usually acquired by measuring at \(N\) sample points \(s_j\) in \([0, 1]\); we take \(s_j = (j-1)/(N-1)\) for \(j = 1, \ldots, N\). Consider \(\mathbb{R}^N\) endowed with the inner product

$$\langle u, v \rangle = \frac{1}{N} \sum_{j=1}^{N} u_j v_j, \quad u = (u_j), v = (v_j) \in \mathbb{R}^N.$$ 

We define \(A : L^2[0, 1] \to \mathbb{R}^N\) by

$$Ax := \left( \int_0^1 k(s, s') x(s') \, ds', \ldots, \int_0^1 k(s_N, s') x(s') \, ds' \right), \quad x \in L^2[0, 1]$$

which is clearly a bounded linear operator. In our numerical simulations, we take \(N = 401\) and add independent Gaussian noise in \(N(0, \sigma^2)\) with \(\sigma = 0.2\) to the exact data \(y := (y(s_1), \ldots, y(s_N))\) to produce independent identically distributed noisy data \(y_i, i = 1, \ldots, n\), with \(E[y_i] = y\). To discretize the problem, we divide \([0, 1]\) into \(m = 400\) subintervals of equal length with nodal points \(s_j = j/m, j = 0, \ldots, m\), and approximate integrals by the trapezoidal rule. For implementing (4.2) and (4.3) with the noisy data \(y_i\) we use the step-size \(\gamma = 2/\|A\|^2\) and terminate the iterations by rule 3.8 with \(\beta_0 = 10\) and \(\tau_n\) given by (3.14), where \(\tau_0 = 1.1\). In figure 1 we plot the exact data (blue one) and the noisy data (10 samples, green circles). We consider several different sample sizes \(n = 10, 10^2, 10^4, 10^5\), each is run by 200 sim-
ulations. In order to visualize the performance, we plot in figure 1 the reconstructed solutions (the mean of 200 simulations), where ‘dgm-NN’ and ‘Landweber’ represent the results obtained by the methods (4.2) and (4.3) respectively. The results indicate that, as the sample size increases, more accurate reconstruction results can be obtained. Since the method (4.2) incorporates the nonnegativity constraint, it produces satisfactory results; while the reconstruction by (4.3) includes undesired negative values.

In table 1 we report the numerical results including the required average number of iterations and the relative error (in average) which is calculated by \(\sqrt{\frac{1}{200} \sum_{i=1}^{200} e_i^2}\) as an approximation of \((\mathbb{E}[[|\hat{x}_k^{(i)} - x|^2]|\|x^1\|^2])^{1/2}\), where \(e_i\) denotes the relative error between the exact solution and the reconstructed solution of the \(i\)th run respectively. We also record the number of the times that \(t_i\) reaches \(\beta_0 n\) to see if rule 3.8 requires the emergency stop. Besides the advantage that the method (4.2) can capture the nonnegativity feature of the sought solutions, table 1 illustrates that (4.2) can be terminated by rule 3.8 with less number of iterations but still can produce reconstructed solutions with smaller relative errors.
Figure 1. The noisy data (10 samples, green circle), the exact data (blue one) and the reconstruction results with different sample sizes of example 4.1.

Table 1. Numerical results for example 4.1, where the sought solution is nonnegative.

| \( n \) | Method   | Iteration numbers (in average) | Relative error | Emergency stops |
|--------|----------|-------------------------------|----------------|-----------------|
| 10     | dgm-NN   | 17                            | 3.9138 \times 10^{-1} | 0               |
|        | Landweber| 17                            | 3.9254 \times 10^{-1} | 0               |
| \( 10^2 \) | dgm-NN   | 115                           | 1.2506 \times 10^{-1} | 0               |
|        | Landweber| 101                           | 1.5405 \times 10^{-1} | 0               |
| \( 10^3 \) | dgm-NN   | 252                           | 7.4015 \times 10^{-2} | 0               |
|        | Landweber| 356                           | 9.7710 \times 10^{-2} | 0               |
| \( 10^4 \) | dgm-NN   | 887                           | 4.3395 \times 10^{-2} | 0               |
|        | Landweber| 1174                          | 6.1918 \times 10^{-2} | 0               |
| \( 10^5 \) | dgm-NN   | 2686                          | 2.2675 \times 10^{-2} | 0               |
|        | Landweber| 4373                          | 3.7267 \times 10^{-2} | 0               |

In figure 2 we present the boxplots of the relative errors given by the methods (4.2) and (4.3) with different sample sizes. On each box, the central mark is the median, the bottle and top edges of the box indicate the 25th and 75th percentiles, the whiskers extend to the most extreme data points the algorithm considers to be not outliers, and the outliers (red crosses) are plotted individually. It is visible that the proposed method is convergent. The red crosses below the blue box imply that the real noise levels have been overestimated. In this case the upper bound \( \beta_{0n} \) of iteration numbers plays the important role of emergency stop. On the other hand, the red crosses above the blue box imply the real noise levels have been underestimated or the semi-convergence phenomenon has been happened already.

Example 4.2. Consider the equation \( Ax = y \), where \( A : L^1(\mathcal{D}) \to \mathcal{V} \) is a bounded linear operator, \( \mathcal{V} \) is a Hilbert space, and \( \mathcal{D} \subset \mathbb{R}^d \) is a bounded domain. Assuming the sought solution is
Figure 2. Boxplots of the relative errors for 200 simulations for example 4.1 with different sample sizes for noisy data corrupted by Gaussian noise \(N(0, \sigma^2)\) with \(\sigma = 0.2\); the left one is for the method (4.2) and the right one is for the method (4.3).

a probability density function, we may find such a solution by considering the convex minimization problem (1.2) with

\[
R(x) := f(x) + \iota_\Delta(x),
\]

(4.7)

where \(\iota_\Delta\) denotes the indicator function of the closed convex set

\[
\Delta := \left\{ x \in L^1(\mathcal{D}) : x \geq 0 \text{ a.e. on } \mathcal{D} \text{ and } \int_\mathcal{D} x = 1 \right\}
\]

in \(L^1(\mathcal{D})\) and \(f\) denotes the negative of the Boltzmann–Shannon entropy, i.e.

\[
f(x) := \begin{cases} \int_\mathcal{D} x \log x & \text{if } x \in L^1_+(\mathcal{D}) \text{ and } x \log x \in L^1(\mathcal{D}), \\ \infty & \text{otherwise,} \end{cases}
\]

where \(L^1_+(\mathcal{D}) := \{ x \in L^1(\mathcal{D}) : x \geq 0 \text{ a.e. on } \mathcal{D} \}\). According to [4], \(R\) satisfies assumption 3.1(ii) with \(c_0 = 1/2\). Furthermore, we can show that for any \(\ell \in L^\infty(\mathcal{D})\) the unique minimizer of

\[
\min_{x \in L^1(\mathcal{D})} \left\{ R(x) - \int_\mathcal{D} \ell x \right\}
\]

is given by \(\hat{x} := e^\ell / \int_\mathcal{D} e^\ell\), see [26]. Therefore the dual gradient method (1.5) using multiple repeated measurement data takes the form

\[
x_t^{(n)} = \frac{1}{\int_\Omega e^{\lambda^{(n)}_{t+1}}} e^{\lambda^{(n)}_{t+1}}, \quad \lambda^{(n)}_{t+1} = \lambda^{(n)}_t - \gamma \left(Ax^{(n)}_t - y^{(n)}\right)
\]

(4.8)

with initial guess \(\lambda_0^{(n)} = 0\), which is an entropic dual gradient method [24].

To test the performance of the method (4.8), we consider again the ill-posed integral equation (4.4) with the kernel \(k(s, s') = 4e^{-\frac{(s-s')^2}{0.0006}}\) and the exact solution

\[
x^\dagger(s) = c \left( e^{\frac{(s-0.2)^2}{0.0006}} + 3e^{\frac{(s-0.7)^2}{0.0006}} \right),
\]

where \(c\) is chosen to ensure \(\int_0^1 x(s) ds = 1\) so that \(x^\dagger\) is a probability density function. For numerical simulations, we use the same procedure in example 4.1 to acquire data and approximate integrals with \(N = m = 400\). We then define \(A\) as in (4.6) which is clearly a bounded linear operator from \(L^1[0,1]\) to \(\mathbb{R}^N\). The independent, identically distributed, unbiased noisy
Inverse Problems 39 (2023) 085005
Q Jin and W Wang

Figure 3. The noisy data (ten samples, green circle), the exact data (blue one) and the reconstruction results with different sample sizes for example 4.2.

Table 2. Numerical results for example 4.2, where the sought solution is a probability density function.

| $n$     | Iteration numbers (in average) | Relative error | Emergency stops |
|---------|--------------------------------|----------------|-----------------|
| 10      | 14                             | $1.4828 \times 10^{-1}$ | 0               |
| $10^2$  | 38                             | $1.0258 \times 10^{-1}$ | 0               |
| $10^3$  | 147                            | $3.8150 \times 10^{-1}$ | 0               |
| $10^4$  | 424                            | $2.5232 \times 10^{-2}$ | 0               |
| $10^5$  | 1799                           | $1.8270 \times 10^{-2}$ | 0               |

data are produced from the exact data by adding Gaussian noise in $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.4$. In figure 3, we plot the exact data (the blue one) and the noisy data (10 samples, green circle).

When implementing the method (4.8), we use the step-size $\gamma = 2/\|A\|^2$ and terminate the iteration by rule 3.8 with $\beta_0 = 10$ and $\tau_n$ given by (3.14) with $\tau_0 = 1.1$. We perform the numerical computation for several different sample sizes $n = 10, 10^2, 10^3, 10^4, 10^5$ each is run by 200 simulations. The reconstruction results in average are plotted in figure 3. In table 2 we report further computational results, including the averaged number of iterations, the $L^1$ relative error in average, and the utilized number of emergence stops. We also present the boxplots of the $L^1$ relative errors for 200 simulations in figure 4. The results indicate that, when sample size increases, the relative error is reduced and the reconstruction accuracy is improved.

Example 4.3. In this example we consider using the method (1.5) to reconstruct sparse solutions for ill-posed problems $Af = u$, where $A : L^2(\mathcal{D}) \to \mathcal{U}$ is a bounded linear operator, $\mathcal{U}$ is a Hilbert space, and $\mathcal{D} \subset \mathbb{R}^d$ is a bounded domain. For this purpose, we take $\mathcal{R}(f)$ to be a strongly convex perturbation of $\|f\|_{L^1(\mathcal{D})}$, i.e.
\[ R(f) = \| f \|_{L^1(\Omega)} + \frac{1}{2\beta} \| f \|_{L^2(\Omega)}^2, \]

where \( \beta > 0 \) is a large number. The method (1.5) then becomes

\[ f^{(n)}_t = \beta \text{sign} (A^* \lambda^{(n)}_t) \max \{|A^* \lambda^{(n)}_t| - 1\}, \]
\[ \lambda^{(n)}_{t+1} = \lambda^{(n)}_t - \gamma (A \hat{u}^{(n)} - \bar{u}^{(n)}), \]

where \( \bar{u}^{(n)} \) is the average of a sequence of unbiased independent identically distributed noisy data \( \{u_i\} \) for \( i = 1, \ldots, n \), of the exact data \( u \).

For numerical simulations we consider determining the initial data \( f \) in the time fractional diffusion equation

\[ \begin{aligned}
\partial^\alpha_t u(x, y; \tau) - \triangle u(x, y; \tau) &= 0, \quad (x, y) \in D, \ \tau > 0, \\
u(x, y; 0) &= f(x, y), \quad (x, y) \in \mathbb{R}^2,
\end{aligned} \]

from the measurement of \( u(x_i, y_j; T) \) at a fixed later time \( T > 0 \) on the grid points \( (x, y_j) := (i/N, j/N), \) \( i, j = 0, 1, \ldots, N \), where \( D = [0, 1] \times [0, 1], 0 < \alpha < 1 \), and \( \partial^\alpha_t u \) denotes the Caputo fractional derivative

\[ \partial^\alpha_t u(x, y; \tau) = \frac{1}{\Gamma(1 - \alpha)} \int_0^\tau \frac{1}{(\tau - s)\alpha} \partial_u (x, y; s) ds \]

with \( \Gamma \) denoting the Gamma function. The mapping from \( f \) to \( (u(x_i, y_j; T)) \) is a bounded linear operator from \( L^2(D) \) to \( \mathbb{R}^{(N+1) \times (N+1)} \), where \( \mathbb{R}^{(N+1) \times (N+1)} \) is endowed with the discrete \( L^2 \)-norm. Time fractional diffusion equations occur naturally in anomalous diffusion in which the variance of the process behaves like a non-integer power of time [37], sharply contrast to the classical normal diffusion which is governed by the heat equation.

To solve this inverse problem numerically, we discretize \( D \) by taking \( (N+1) \times (N+1) \) grid points \( (x, y_j) := (i/N, j/N), \) \( i, j = 0, 1, \ldots, N \) and write \( u_{ij}(\tau) \) for \( u(x_i, y_j; \tau) \) and \( f_{ij} \) for \( f(x_i, y_j) \).

Let \( h = 1/N \). Then, by the finite difference approximation of \( -\triangle u \), the diffusion equation becomes

\[ \begin{align}
\text{Figure 4. Boxplots of the relative errors for 200 simulations with different sample sizes for example 4.2.}
\end{align} \]
\[ \partial_t^\alpha u_{ij} + \frac{1}{H^2} (4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}) = 0, \]
\[ i,j = 1, \ldots, N - 1, \]
\[ u_{0,j} = u_{N,j} = u_{i,0} = u_{i,N} = 0, \quad i,j = 0, \ldots, N, \]
\[ u_{ij}(0) = f_{ij}, \quad i,j = 1, \ldots, N - 1. \]

It turns out that the solution of (4.10) has the form
\[ u_{ij}(\tau) = \sum_{p,q=1}^{N-1} c_{p,q}(\tau) \sin(iph\pi) \sin(qj\phi\pi), \]
where each \( c_{p,q}(\tau) \) satisfies the fractional ordinary differential equation
\[ \partial_t^\alpha c_{p,q} + \mu_{p,q} c_{p,q} = 0, \quad \text{with} \quad \mu_{p,q} = \frac{1}{H^2} (4 - 2\cos(p\phi\pi) - 2\cos(q\phi\pi)). \]

Let \( S \) and \( S^{-1} \) denote the discrete sine transform and the inverse discrete sine transform defined respectively by
\[ (Sv)_{p,q} := 4h^2 \sum_{i,j=1}^{N-1} v_{ij} \sin(iph\pi) \sin(qj\phi\pi), \]
\[ (S^{-1}v)_{i,j} := \sum_{p,q=1}^{N-1} v_{p,q} \sin(iph\pi) \sin(qj\phi\pi) \]
for any \((N-1) \times (N-1)\) matrix \((v_{ij})\). Then we have \( c_{p,q}(0) = (Sf)_{p,q} \) and thus \( c_{p,q}(\tau) = (Sf)_{p,q} E_\alpha(-\mu_{p,q} \tau^\alpha) \), where \( E_\alpha \) denotes the Mittag–Leffler function
\[ E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}. \]
Consequently
\[ u_{ij}(\tau) = \sum_{p,q=1}^{N-1} (Sf)_{p,q} E_\alpha(-\mu_{p,q} \tau^\alpha) \sin(iph\pi) \sin(qj\phi\pi) \]
for \( i,j = 1, \ldots, N - 1 \). If we define the transform \( W \) by \( (Wv)_{p,q} = E_\alpha(-\mu_{p,q} \tau^\alpha) v_{p,q} \), then \( u(\tau) = S^{-1} W S f \). Let \( A = S^{-1} W S \). Then \( f \) can be determined by solving \( Af = u \), where \( u := (u_{ij}(T)) \).

In our numerical simulation, we assume the sought solution is sparse and consider \( D = [0, 1] \times [0, 1] \) on an equidistant grid of \( 256 \times 256 \) points. The sought solution \( f \) and the corresponding exact data \( u \) are plotted in figure 5 (the left and middle ones). To reconstruct \( f \), we use multiple repeated measurement data of different sample sizes which are generated from \( u \) by adding independent Gaussian noise \( N(0, \sigma^2) \) with \( \sigma = 0.2 \max |u| \); one sample of measurement data is plotted in figure 5 (the right one).

When applying the method (4.9), we take \( \beta = 300 \) and use the step-size \( \gamma = 2/(\beta \|A\|^2) \). The iteration is terminated by rule 3.8 with \( \beta_0 = 50 \) and \( \tau_n \) given by (3.14); where \( \tau_0 = 1.1 \). The reconstruction results for several different sample sizes \( n = 10, 10^2, 10^3, 10^4 \) are plotted in figure 6, each is based on the average of 200 runs. In table 3 we report the average number of iterations, the average relative errors, and the number of emergency stops. Furthermore, we present the boxplots of the relative errors in figure 7. All these results demonstrate that the
Figure 5. The exact solution, the exact data and the noisy data (random sample) for example 4.3.

Figure 6. The reconstruction results of the method (4.9) with different sample sizes for example 4.3.

The proposed method can give satisfactory reconstruction results when the sample size increases and the sparsity of the sought solution can be captured well.

Example 4.4. In this final example we consider using the method (1.5) to reconstruct piecewise constant solutions. We consider again the equation (4.4) with the kernel given by (4.5) and assume that the sought solution $x^\dagger$ is piecewise constant. By dividing $[0, 1]$ into $m = 400$
subintervals of equal length and approximating integrals by the trapezoidal rule, we have a discrete ill-posed problem $\mathbf{A} \mathbf{x} = \mathbf{y}$, where $\mathbf{A}$ is a matrix. We use the model

$$
\min \left\{ R(x) := \|Dx\|_1 + \frac{1}{2\beta} \|Dx\|_2^2 + \frac{1}{2\beta} \|x\|_2^2 : \mathbf{A} \mathbf{x} = \mathbf{y} \right\},
$$

(4.11)

where $\mathbf{D}$ denotes the discrete gradient operator and $\beta$ is a large positive number. Thus, $R$ is a strongly convex perturbation of the total variation $\|Dx\|_1$. If we apply the method (1.5) to (4.11) directly, we need to solve a minimization problem related to $R$ to obtain $\mathbf{x}^{(n)}$ at each iteration. This can make the algorithm time-consuming since those minimization problems cannot be solved explicitly.

To circumvent this difficulty, by introducing $\mathbf{z} = \mathbf{Dx}$ we reformulate (4.11) as

$$
\min \left\{ f(\mathbf{z}) + \phi(\mathbf{x}) : \mathbf{A} \mathbf{x} = \mathbf{y} \text{ and } \mathbf{Dx} - \mathbf{z} = 0 \right\},
$$

(4.12)

where

$$
f(\mathbf{z}) = \|\mathbf{z}\|_1 + \frac{1}{2\beta} \|\mathbf{z}\|_2^2, \quad \phi(\mathbf{x}) = \frac{1}{2\beta} \|\mathbf{x}\|_2^2.
$$

Since $f(\mathbf{z}) + \phi(\mathbf{x})$ is strongly convex, we may apply the method (1.5) to (4.12) to obtain the iteration scheme

$$
\begin{align*}
\left( \mathbf{x}^{(n)}_t, \mathbf{z}^{(n)}_t \right) &= \arg \min_{\mathbf{x}, \mathbf{z}} \left\{ f(\mathbf{z}) + \phi(\mathbf{x}) - \langle \lambda^{(n)}_t, \mathbf{Ax} \rangle - \langle \mu^{(n)}_t, \mathbf{Dx} - \mathbf{z} \rangle \right\}, \\
\left( \lambda^{(n)}_{t+1}, \mu^{(n)}_{t+1} \right) &= \left( \lambda^{(n)}_t, \mu^{(n)}_t \right) - \gamma \left( \mathbf{A} \mathbf{x}^{(n)}_t - \bar{\mathbf{y}}^{(n)}, \mathbf{Dx}^{(n)}_t - \mathbf{z}^{(n)} \right).
\end{align*}
$$

---

**Table 3.** Numerical results for example 4.3, where the sought solution is sparse.

| n   | Iteration numbers (in average) | Relative error | Emergency stops |
|-----|-------------------------------|----------------|----------------|
| 10  | 174                           | $5.9865 \times 10^{-1}$ | 0              |
| $10^2$ | 458                           | $3.8077 \times 10^{-1}$ | 0              |
| $10^3$ | 933                           | $1.7786 \times 10^{-1}$ | 0              |
| $10^4$ | 1821                          | $1.1831 \times 10^{-1}$ | 0              |

**Figure 7.** Boxplots of the relative errors for 200 simulations with different sample sizes for example 4.3.
Figure 8. The noisy data (10 samples, green circle), the exact data (blue one) and the reconstruction results with different sample sizes for example 4.4.

Since $x_t^{(n)}$ and $z_t^{(n)}$ can be given explicitly, this leads to the following algorithm

\[
\begin{align*}
x_t^{(n)} &= \beta \left( A^T \lambda_t^{(n)} + D^T \mu_t^{(n)} \right), \\
z_t^{(n)} &= -\beta \text{sign} \left( \mu_t^{(n)} \right) \max \left\{ \left| \mu_t^{(n)} \right| - 1, 0 \right\}, \\
\lambda_{t+1}^{(n)} &= \lambda_t^{(n)} - \gamma \left( A x_t^{(n)} - \bar{y}^{(n)} \right), \\
\mu_{t+1}^{(n)} &= \mu_t^{(n)} - \gamma \left( D x_t^{(n)} - z_t^{(n)} \right).
\end{align*}
\] (4.13)

In our numerical experiments, the sought solution is piecewise constant, whose graph together with the exact data is plotted in figure 8 in blue. In order to apply the method (4.13), we use multiple repeated measurement data of several different sample sizes $n = 10, 10^2, 10^3, 10^4, 10^5$ corrupted by independent Gaussian noise $N(0, \sigma^2)$ with $\sigma = 0.2$ and let $\bar{y}^{(n)}$ be the average of these data; 10 samples of these measurement data are plotted in figure 8 in green. The step-size in (4.13) is chosen to be $\gamma = 2/|\beta||A|^2 + 4|$. In order to enhance the effect of total variation in reconstruction, we take $\beta = 100$. The method is terminated by rule 3.8 adapted to (4.12) with $\beta_0 = 200$ and $\tau_n$ given by (3.14), where $\tau_0 = 1.1$. For each sample size, we run 200 simulations and the reconstruction results in average are plotted in figure 8 in red. In table 4 we report further numerical results including the average number of iterations, the relative error in average, and the number of emergency stops used. The boxplot of the relative error is presented in figure 9. These results clearly demonstrate that the proposed method converges and more and more satisfactory reconstruction results can be obtained when the sample size increases.
Table 4. Numerical results for example 4.4, where the sought solution is piecewise constant.

| $n$    | Iteration numbers (in average) | Relative error | Emergency stops |
|--------|--------------------------------|----------------|-----------------|
| $10^{1}$ | 1109                           | $3.5812 \times 10^{-1}$ | 10              |
| $10^{2}$ | 17 877                         | $2.2729 \times 10^{-1}$ | 70              |
| $10^{3}$ | 64 078                         | $1.7208 \times 10^{-1}$ | 0               |
| $10^{4}$ | 242 623                        | $1.3263 \times 10^{-1}$ | 0               |
| $10^{5}$ | 675 068                        | $1.0761 \times 10^{-1}$ | 0               |

Figure 9. Boxplots of the relative errors for 200 simulations with different sample sizes for example 4.4.

Data availability statement

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

Acknowledgment

The work of Q Jin is partially supported by the Future Fellowship of the Australian Research Council (FT170100231). The work of W Wang is supported by the National Natural Science Foundation of China (No. 12071184) and the Top-level Talent Project of Zhejiang Province.

ORCID iDs

Qinian Jin https://orcid.org/0000-0001-9283-9791
Wei Wang https://orcid.org/0000-0002-3316-8438

References

[1] Bakushinskii A B 1984 Remarks on choosing a regularization parameter using the quasioptimality and ratio criterion USSR Comput. Math. Math. Phys. 24 181–2
[2] Blanchard G and Mathé P 2012 Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration Inverse Problems 28 115011

[3] Beck A and Teboulle M 2014 A fast dual proximal gradient algorithm for convex minimization and applications Oper. Res. Lett. 42 1–6

[4] Borwein J M and Lewis C S 1991 Convergence of best entropy estimates SIAM J. Optim. 1 191–205

[5] Bürj 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

[6] Chen D H, Jiang D, Yousept I and Zou J 2022 Variational source conditions for inverse Robin and flux problems by partial measurements Inverse Problems Imaging 16 283–304

[7] Chen D H and Yousept I 2019 Variational source condition for ill-posed backward nonlinear Maxwell’s equations Inverse Problems 35 025001

[8] Engh W, Hanke M and Neubauer A 1996 Regularization of Inverse Problems (Dordrecht: Kluwer)

[9] Flemming J 2018 Existence of variational source conditions for nonlinear inverse problems in Banach spaces J. Inverse Ill-Posed Problems 26 277–86

[10] Frick K and Scherzer O 2010 Regularization of ill-posed linear equations by the non-stationary augmented Lagrangian method J. Integral Equ. Appl. 22 277–86

[11] Grasmair M 2020 Source conditions for non-quadratic Tikhonov regularization Numer. Funct. Anal. Optim. 41 1352–72

[12] Hansen P C and Leary D P O 1993 The use of the L-curve in the regularization of discrete ill-posed problems SIAM J. Sci. Comput. 14 1487–503

[13] Harrach B, Jahn T and Potthast R 2020 Beyond the Bakushinskii veto: regularising linear inverse problems without knowing the noise distribution Numer. Math. 145 581–603

[14] Harrach B, Jahn T and Potthast R 2023 Regularizing linear inverse problems under unknown non-Gaussian white noise allowing repeated measurements IMA J. Numer. Anal. 43 443–500

[15] Hassan U and Anwar M S 2010 Reducing noise by repetition: introduction to signal averaging Eur. J. Phys. 31 453–65

[16] Hofmann B, Kaltenbacher B, Pöschl C and Scherzer O 2007 A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators Inverse Problems 23 987–1010

[17] Hofmann B and Mathé P 2012 Parameter choice under variational inequalities Inverse Problems 28 104006

[18] Hofhage T and Weidling F 2015 Verification of a variational source condition for acoustic inverse medium scattering problems Inverse Problems 31 075006

[19] Weidling F and Hofhage T 2017 Variational source condition and stability estimates for inverse electromagnetic medium scattering problems Inverse Problems Imaging 11 203–20

[20] Hofhage T and Weidling F 2017 Characterizations of variational source conditions, converse results and maxisets of spectral regularization methods SIAM J. Numer. Anal. 55 598–620

[21] Jahn T 2021 A modified discrepancy principle to attain optimal convergence rates under unknown noise Inverse Problems 37 095008

[22] Jin Q 2016 Hanke-Raus heuristic rule for variational regularization in Banach spaces Inverse Problems 32 085008

[23] Jin Q 2017 On a heuristic stopping rule for the regularization of inverse problems by the augmented Lagrangian method Numer. Math. 136 973–92

[24] Jin Q 2022 Convergence rates of a dual gradient method for constrained linear ill-posed problems Numer. Math. 151 841–71

[25] Jin Q and Amato U 2001 A discrete scheme of Landweber iteration for solving nonlinear ill-posed problems J. Math. Anal. Appl. 253 187–203

[26] Jin Q, Lu X and Zhang L 2023 Stochastic mirror descent method for linear ill-posed problems in Banach spaces Inverse Problems 39 095008

[27] Jin Q and Mathé P 2013 Oracle inequality for a statistical Raus–Gfrerer-type rule Int. J. Uncertain. Quantification 1 386–407

[28] Jin Q and Stals L 2012 Nonstationary iterated Tikhonov regularization for ill-posed problems in Banach spaces Inverse Problems 28 104011

[29] Jin Q and Wang W 2013 Landweber iteration of Kaczmarz type with general non-smooth convex penalty functionals Inverse Problems 29 085011

[30] Kindermann S and Neubauer A 2008 On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization Inverse Problems Imaging 2 291–9
[31] Kindermann S and Raik K 2020 Convergence of heuristic parameter choice rules for convex Tikhonov regularization *SIAM J. Numer. Anal.* **58** 1773–800
[32] Ledoux M and Talagrand M 1991 *Probability in Banach Spaces: Isoperimetry and Processes* vol 23 (Berlin: Springer)
[33] Lu S and Mathé P 2014 Discrepancy based model selection in statistical inverse problems *J. Complexity* **30** 290–308
[34] Lyons R G 2004 *Understanding Digital Signal Processing* vol 3/E (Chennai: Pearson Education India)
[35] Nemirovski S and Yudin D B 1983 *Problem Complexity and Method Efficiency in Optimization* (New York: Wiley)
[36] Plato R and Vainikko G 1990 On the regularization of projection methods for solving ill-posed problems *Numer. Math.* **57** 63–79
[37] Podlubny I 1999 *Fractional Differential Equations* (San Diego, CA: Academic)
[38] 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) (Berlin: Walter de Gruyter)
[39] Tseng P 1991 Applications of a splitting algorithm to decomposition in convex programming and variational inequalities *SIAM J. Control Optim.* **29** 119–38
[40] Uzawa H 1958 Iterative methods for concave programming *Studies in Linear and Nonlinear Programming* (Cambridge: Cambridge University Press) pp 154–65
[41] van Drongelen W 2018 *Signal Processing for Neuroscientists* 2nd edn (New York: Academic)
[42] Zălinescu C 2002 *Convex Analysis in General Vector Spaces* (River Edge, NJ: World Scientific Publishing Co., Inc.)