# Convergence rates of a dual gradient method for constrained linear ill-posed problems

Qinian Jin[1]

## Abstract

In this paper we consider a dual gradient method for solving linear ill-posed problems $Ax = y$, where $A : X \to Y$ is a bounded linear operator from a Banach space $X$ to a Hilbert space $Y$. A strongly convex penalty function is used in the method to select a solution with desired feature. Under variational source conditions on the sought solution, convergence rates are derived when the method is terminated by either an *a priori* stopping rule or the discrepancy principle. We also consider an acceleration of the method as well as its various applications.

**Mathematics Subject Classification** 65J15 · 65J20 · 47H17

## 1 Introduction

Many linear inverse problems can be formulated into the minimization problem

$$\min\{\mathcal{R}(x) : x \in X \text{ and } Ax = y\}, \tag{1.1}$$

where $A : X \to Y$ is a bounded linear operator from a Banach space $X$ to a Hilbert space $Y$, $y \in \text{Ran}(A)$, the range of $A$, and $\mathcal{R} : X \to (-\infty, \infty]$ is a proper, lower semi-continuous, convex function that is used to select a solution with desired feature. Throughout the paper, all spaces are assumed to be real vector spaces; however, all results still hold for complex vector spaces by minor modifications adapted to complex environments. The norms in $X$ and $Y$ are denoted by the same notation $\|\cdot\|$. We also use the same notation $\langle\cdot,\cdot\rangle$ to denote the duality pairing in Banach spaces and the inner product in Hilbert spaces. When the operator $A$ does not have a closed range, the problem (1.1) is ill-posed in general, thus, if instead of $y$, we only have a noisy data $y^\delta$ satisfying

✉ Qinian Jin
Qinian.Jin@anu.edu.au

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

$$\|y^\delta - y\| \leq \delta$$

with a small noise level $\delta > 0$, then replacing $y$ in (1.1) by $y^\delta$ may lead to a problem that is not well-defined; even if it is well-defined, the solution may not depend continuously on the data. In order to use a noisy data to find an approximate solution of (1.1), a regularization technique should be employed to remove the instability [13, 42]

In this paper we will consider a dual gradient method to solve (1.1). This method is based on applying the gradient method to its dual problem. In order for a better understanding, we provide a brief derivation of this method which is well known in optimization community [5, 43]; the facts from convex analysis that are used will be reviewed in Sect. 2. Assume that we only have a noisy data $y^\delta$ and consider the problem (1.1) with $y$ replaced by $y^\delta$. The associated Lagrangian function is

$$\mathcal{L}(x, \lambda) = \mathcal{R}(x) - \langle \lambda, Ax - y^\delta \rangle, \quad x \in X \text{ and } \lambda \in Y$$

which induces the dual function

$$\inf_{x \in X} \left\{ \mathcal{R}(x) - \langle \lambda, Ax - y^\delta \rangle \right\} = -\mathcal{R}^*(A^*\lambda) + \langle \lambda, y^\delta \rangle,$$

where $A^* : Y \to X^*$ denotes the adjoint of $A$ and $\mathcal{R}^* : X^* \to (-\infty, \infty]$ denotes the Legendre-Fenchel conjugate of $\mathcal{R}$. Thus the corresponding dual problem is

$$\min_{\lambda \in Y} \left\{ d_{y^\delta}(\lambda) := \mathcal{R}^*(A^*\lambda) - \langle \lambda, y^\delta \rangle \right\}. \tag{1.2}$$

Assuming that $\mathcal{R}$ is strongly convex, then $\mathcal{R}^*$ is continuous differentiable with $\nabla \mathcal{R}^* : X^* \to X$ and so is the function $\lambda \to d_{y^\delta}(\lambda)$ on $Y$. Therefore, we may apply a gradient method to solve (1.2) which leads to

$$\lambda_{n+1} = \lambda_n - \gamma \left( A\nabla \mathcal{R}^*(A^*\lambda_n) - y^\delta \right),$$

where $\gamma > 0$ is a step-size. Let $x_n := \nabla \mathcal{R}^*(A^*\lambda_n)$. Then by the properties of subdifferential we have $A^*\lambda_n \in \partial \mathcal{R}(x_n)$ and hence

$$x_n \in \arg\min_{x \in X} \left\{ \mathcal{R}(x) - \langle \lambda_n, Ax - y^\delta \rangle \right\}.$$

Combining the above two equations results in the following dual gradient method

$$\begin{aligned}
x_n &= \arg\min_{x \in X} \left\{ \mathcal{R}(x) - \langle \lambda_n, Ax - y^\delta \rangle \right\}, \\
\lambda_{n+1} &= \lambda_n - \gamma (Ax_n - y^\delta).
\end{aligned} \tag{1.3}$$

Note that when $X$ is a Hilbert space and $\mathcal{R}(x) = \|x\|^2/2$, the method (1.3) becomes the standard linear Landweber iteration in Hilbert spaces [13].

By setting $\xi_n := A^*\lambda_n$, we can obtain from (1.3) the algorithm

$$
\begin{aligned}
x_n &= \arg\min_{x \in X} \left\{ \mathcal{R}(x) - \langle \xi_n, x \rangle \right\}, \\
\xi_{n+1} &= \xi_n - \gamma A^* \left( Ax_n - y^\delta \right).
\end{aligned}
\tag{1.4}
$$

Actually the method (1.4) is equivalent to (1.3) when the initial guess $\xi_0$ is chosen from Ran($A^*$), the range of $A^*$. Indeed, under the given condition on $\xi_0$, we can conclude from (1.4) that $\xi_n \in$ Ran($A^*$) for all $n$. Assuming $\xi_n = A^*\lambda_n$ for some $\lambda_n \in Y$, we can easily see that $x_n$ defined by the first equation in (1.4) satisfies the first equation in (1.3). Furthermore, from the second equation in (1.4) we have

$$
\xi_{n+1} = A^* \left( \lambda_n - \gamma(Ax_n - y^\delta) \right)
$$

which means $\xi_{n+1} = A^*\lambda_{n+1}$ with $\lambda_{n+1}$ defined by the second equation in (1.3).

The method (1.4) as well as its generalizations to linear and nonlinear ill-posed problems in Banach spaces have been considered in [9, 29, 33, 34, 40, 41] and the convergence property has been proved when the method is terminated by the discrepancy principle. However, except for the linear and nonlinear Landweber iteration in Hilbert spaces [13, 22], the convergence rate in general is missing from the existing convergence theory. In this paper we will consider the dual gradient method (1.3) and hence the method (1.4) under the discrepancy principle

$$
\|Ax_{n_\delta} - y^\delta\| \le \tau\delta < \|Ax_n - y^\delta\|, \quad 0 \le n < n_\delta
\tag{1.5}
$$

with a constant $\tau > 1$ and derive the convergence rate when the sought solution satisfies a variational source condition. This is the main contribution of the present paper. We also consider accelerating the dual gradient method by Nesterov's acceleration strategy and provide a convergence rate result when the method is terminated by an *a priori* stopping rule. Furthermore, we discuss various applications of our convergence theory: we provide a rather complete analysis of the dual projected Landweber iteration for solving linear ill-posed problems in Hilbert spaces with convex constraint which was proposed in [12] with only preliminary results; we also propose an entropic dual gradient method using Boltzmann-Shannon entropy to solve linear ill-posed problems whose solutions are probability density functions.

In the existing literature there exist a number of regularization methods for solving (1.1), including the Tikhonov regularization method, the augmented Lagrangian method, and the nonstationary iterated Tikhonov regularization [19, 32, 42]. In particular, we would like to mention that the augmented Lagrangian method

$$
\begin{aligned}
x_n &\in \arg\min_{x \in X} \left\{ \mathcal{R}(x) - \langle \lambda_n, Ax - y^\delta \rangle + \frac{\gamma_n}{2} \|Ax - y^\delta\|^2 \right\}, \\
\lambda_{n+1} &= \lambda_n - \gamma_n(Ax_n - y^\delta)
\end{aligned}
\tag{1.6}
$$

has been considered in [17–19, 30] for solving ill-posed problem (1.1) as a regularization method. This method can be viewed as a modification of the dual gradient

method (1.3) by adding the augmented term $\frac{\gamma_n}{2}\|Ax - y^\delta\|^2$ to the definition of $x_n$. Although the addition of this extra term enables to establish the regularization property of the augmented Lagrangian method under quite general conditions on $\mathcal{R}$, it destroys the decomposability structure and thus extra work has to be done to determine $x_n$ at each iteration step. In contrast, the convergence analysis of the dual gradient method requires $\mathcal{R}$ to be strongly convex, however the determination of $x_n$ is much easier in general. In fact $x_n$ can be given by a closed formula in many interesting cases; even if $x_n$ does not have a closed formula, there exist fast algorithms for solving the minimization problem that is used to define $x_n$ since it does not involve the operator $A$, see Sect. 4 and [9, 29, 33] for instance. This can significantly save the computational time.

The paper is organized as follows, In Sect. 2, we give a brief review of some basic facts from convex analysis in Banach spaces. In Sect. 3, after a quick account on convergence, we focus on deriving the convergence rates of the dual gradient method under variational source conditions on the sought solution when the method is terminated by either an *a priori* stopping rule or the discrepancy principle; we also discuss the acceleration of the method by Nesterov's strategy. Finally in Sect. 4, we address various applications of our convergence theory.

## 2 Preliminaries

In this section, we will collect some basic facts on convex analysis in Banach spaces which will be used in the analysis of the dual gradient method (1.3); for more details one may refer to [8, 44] for instance.

Let $X$ be a Banach space whose norm is denoted by $\|\cdot\|$, we use $X^*$ to denote its dual space. Given $x \in X$ and $\xi \in X^*$ we write $\langle \xi, x \rangle = \xi(x)$ for the duality pairing. For a convex function $f : X \to (-\infty, \infty]$, we use

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

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

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

The collection of all subgradients of $f$ at $x$ is denoted as $\partial f(x)$ and is called the subdifferential of $f$ at $x$. If $\partial f(x) \neq \emptyset$, then $f$ is called subdifferentiable at $x$. Thus $x \to \partial f(x)$ defines a set-valued mapping $\partial f$ whose domain of definition is defined as

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

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

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

which is always nonnegative.

For a proper function $f : X \to (-\infty, \infty]$, its Legendre–Fenchel conjugate is defined by

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

which is a convex function taking values in $(-\infty, \infty]$. According to the definition we immediately have the Fenchel–Young inequality

$$f^*(\xi) + f(x) \geq \langle \xi, x \rangle \tag{2.1}$$

for all $x \in X$ and $\xi \in X^*$. If $f : X \to (-\infty, \infty]$ is proper, lower semi-continuous and convex, $f^*$ is also proper and

$$\xi \in \partial f(x) \Longleftrightarrow x \in \partial f^*(\xi) \Longleftrightarrow f(x) + f^*(\xi) = \langle \xi, x \rangle. \tag{2.2}$$

We will use the following version of the Fenchel–Rockafellar duality formula (see [8, Theorem 4.4.3]).

**Proposition 2.1** *Let $X$ and $Y$ be Banach spaces, let $f : X \to (-\infty, \infty]$ and $g : Y \to (-\infty, \infty]$ be proper, convex functions, and let $A : X \to Y$ 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*

$$\inf_{x \in X} \{ f(x) + g(Ax) \} = \sup_{\eta \in Y^*} \{ -f^*(A^*\eta) - g^*(-\eta) \}. \tag{2.3}$$

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

$$f(t\bar{x} + (1-t)x) + \sigma t(1-t)\|\bar{x} - x\|^2 \leq t f(\bar{x}) + (1-t) f(x) \tag{2.4}$$

for all $\bar{x}, x \in \text{dom}(f)$ and $t \in [0, 1]$. The largest number $\sigma > 0$ such that (2.4) holds true is called the modulus of convexity of $f$. It can be shown that a proper, lower semi-continuous, convex function $f : X \to (-\infty, \infty]$ is strongly convex with modulus of convexity $\sigma > 0$ if and only if

$$D_f^\xi(\bar{x}, x) \geq \sigma \|x - \bar{x}\|^2 \tag{2.5}$$

for all $\bar{x} \in \text{dom}(f)$, $x \in \text{dom}(\partial f)$ and $\xi \in \partial f(x)$; see [44, Corollary 3.5.11]. Furthermore, [44, Corollary 3.5.11] also contains the following important result which in

particular shows that the strong convexity of $f$ implies the continuous differentiability of $f^*$.

**Proposition 2.2** *Let $X$ be a Banach space and let $f : X \to (-\infty, \infty]$ be a proper, lower semi-continuous, strongly convex function with modulus of convexity $\sigma > 0$. Then $dom(f^*) = X^*$, $f^*$ is Fréchet differentiable and its gradient $\nabla f^* : X^* \to X$ satisfies*

$$\|\nabla f^*(\xi) - \nabla f^*(\eta)\| \leq \frac{\|\xi - \eta\|}{2\sigma}$$

*for all $\xi, \eta \in X^*$.*

It should be emphasized that $X$ in Proposition 2.2 can be an arbitrary Banach space. For the gradient $\nabla f^*$ of $f^*$, it is in general a mapping from $X^* \to X^{**}$, the second dual space of $X$. Proposition 2.2 actually concludes that, for each $\xi \in X^*$, $\nabla f^*(\xi)$ is an element in $X^{**}$ that can be identified with an element in $X$ via the canonical embedding $X \to X^{**}$, and thus $\nabla f^*$ is a mapping from $X^*$ to $X$.

## 3 Main results

This section focuses on the study of the dual gradient method (1.3). We will make the following assumption.

**Assumption 1** (i) $X$ is a Banach space, $Y$ is a Hilbert space, and $A : X \to Y$ is a bounded linear operator;
(ii) $\mathcal{R} : X \to (-\infty, \infty]$ is a proper, lower semi-continuous, strongly convex function with modulus of convexity $\sigma > 0$;
(iii) The equation $Ax = y$ has a solution in $dom(\mathcal{R})$.

Under Assumption 1, one can use [44, Proposition 3.5.8] to conclude that (1.1) has a unique solution $x^\dagger$ and, for each $n$, the minimization problem involved in the method (1.3) has a unique minimizer $x_n$ and thus the method (1.3) is well-defined. By the definition of $x_n$ we have

$$A^*\lambda_n \in \partial\mathcal{R}(x_n). \tag{3.1}$$

By virtue of (2.2) and Proposition 2.2 we further have

$$x_n = \nabla\mathcal{R}^*(A^*\lambda_n) \tag{3.2}$$

for all $n \geq 0$.

### 3.1 Convergence

The regularization property of a family of gradient type methods, including (1.4) as a special case, have been considered in [33] for solving ill-posed problems in Banach

spaces. Adapting the corresponding result to the dual gradient method (1.3) we can obtain the following convergence result.

**Theorem 3.1** *Let Assumption* 1 *hold and let* $L := \|A\|^2/(2\sigma)$. *Consider the dual gradient method* (1.3) *with* $\lambda_0 = 0$ *for solving* (1.1).

(i) *If* $0 < \gamma \le 1/L$ *then for the integer* $n_\delta$ *chosen such that* $n_\delta \to \infty$ *and* $\delta^2 n_\delta \to 0$ *as* $\delta \to 0$ *there hold*

$$\mathcal{R}(x_{n_\delta}) \to \mathcal{R}(x^\dagger) \quad and \quad D_{\mathcal{R}}^{A^*\lambda_{n_\delta}}(x^\dagger, x_{n_\delta}) \to 0$$

*and hence* $\|x_{n_\delta} - x^\dagger\| \to 0$ *as* $\delta \to 0$.

(ii) *If* $\tau > 1$ *and* $\gamma > 0$ *are chosen such that* $1 - 1/\tau - L\gamma > 0$, *then the discrepancy principle* (1.5) *defines a finite integer* $n_\delta$ *with*

$$\mathcal{R}(x_{n_\delta}) \to \mathcal{R}(x^\dagger) \quad and \quad D_{\mathcal{R}}^{A^*\lambda_{n_\delta}}(x^\dagger, x_{n_\delta}) \to 0$$

*and hence* $\|x_{n_\delta} - x^\dagger\| \to 0$ *as* $\delta \to 0$.

Theorem 3.1 gives the convergence results on the method (1.3) with $\lambda_0 = 0$. The convergence result actually holds for any initial guess $\lambda_0$ with the iterative sequence defined by (1.3) converging to a solution $x^\dagger$ of $Ax = y$ with the property

$$D_{\mathcal{R}}^{A^*\lambda_0}(x^\dagger, x_0) = \min\left\{ D_{\mathcal{R}}^{A^*\lambda_0}(x, x_0) : x \in X \text{ and } Ax = y \right\},$$

where $x_0 = \arg\min_{x \in X}\{\mathcal{R}(x) - \langle \lambda_0, x \rangle\}$; this can be seen from Theorem 3.1 by replacing $\mathcal{R}(x)$ by $D_{\mathcal{R}}^{A^*\lambda_0}(x, x_0)$. This same remark applies to the convergence rate results in the forthcoming subsection. For simplicity of exposition, in the following we will consider only the method (1.3) with $\lambda_0 = 0$.

In [33] the convergence result was stated for $X$ to be a reflexive Banach space. The reflexivity of $X$ was only used in [33] to show the well-definedness of each $x_n$ by the procedure of extracting a weakly convergent subsequence from a bounded sequence. Under Assumption 1 (ii) the reflexivity of $X$ is unnecessary as the strong convexity of $\mathcal{R}$ guarantees that each $x_n$ is well-defined in an arbitrary Banach space, see [44, Proposition 3.5.8]. This relaxation on $X$ allows the convergence result to be used in a wider range of applications, see Sect. 4.2 for instance.

The work in [33] actually concentrates on proving part (ii) of Theorem 3.1, i.e. the regularization property of the method terminated by the discrepancy principle and part (i) was not explicitly stated. However, the argument can be easily adapted to obtain part (i) of Theorem 3.1, i.e. the regularization property of the method under an a priori stopping rule.

It should be mentioned that the convergence $\mathcal{R}(x_{n_\delta}) \to \mathcal{R}(x^\dagger)$ was not established in [33]. However, if the residual $\|Ax_n - y^\delta\|$ is monotonically decreasing with respect to $n$, then, following the proof in [33], one can easily establish the convergence $\mathcal{R}(x_{n_\delta}) \to \mathcal{R}(x^\dagger)$ as $\delta \to 0$. For the dual gradient method (1.3), the monotonicity of $\|Ax_n - y^\delta\|$ is established in the following result which is also useful in the forthcoming analysis on deriving convergence rates.

**Lemma 3.2** *Let Assumption* 1 *hold and let* $0 < \gamma \le 4\sigma/\|A\|^2$. *Then for the sequence* $\{x_n\}$ *defined by* (1.3) *there holds*

$$\|Ax_{n+1} - y^\delta\| \le \|Ax_n - y^\delta\|$$

*for all integers* $n \ge 0$.

**Proof** Recall from (3.1) that $A^*\lambda_n \in \partial\mathcal{R}(x_n)$ for each $n \ge 0$. By using (2.5) and the equation $\lambda_{n+1} = \lambda_n - \gamma(Ax_n - y^\delta)$, we have

$$
\begin{aligned}
2\sigma\|x_{n+1} - x_n\|^2 &\le D_\mathcal{R}^{A^*\lambda_n}(x_{n+1}, x_n) + D_\mathcal{R}^{A^*\lambda_{n+1}}(x_n, x_{n+1}) \\
&= \langle A^*\lambda_{n+1} - A^*\lambda_n, x_{n+1} - x_n \rangle \\
&= \langle \lambda_{n+1} - \lambda_n, Ax_{n+1} - Ax_n \rangle \\
&= \gamma \langle Ax_n - y^\delta, Ax_n - Ax_{n+1} \rangle.
\end{aligned}
$$

In view of the polarization identity in Hilbert spaces, we further have

$$
\begin{aligned}
2\sigma\|x_{n+1} - x_n\|^2 &\le \frac{\gamma}{2}\left(\|Ax_n - y^\delta\|^2 - \|Ax_{n+1} - y^\delta\|^2 + \|A(x_{n+1} - x_n)\|^2\right) \\
&\le \frac{\gamma}{2}\left(\|Ax_n - y^\delta\|^2 - \|Ax_{n+1} - y^\delta\|^2\right) + \frac{\gamma\|A\|^2}{2}\|x_{n+1} - x_n\|^2.
\end{aligned}
$$

Since $0 < \gamma \le 4\sigma/\|A\|^2$, we thus obtain the monotonicity of $\|Ax_n - y^\delta\|^2$ with respect to $n$. $\qquad\square$

### 3.2 Convergence rates

In this subsection we will derive the convergence rates of the dual gradient method (1.3) when the sought solution satisfies certain variational source conditions. The following result plays a crucial role for achieving this purpose.

**Proposition 3.3** *Let Assumption* 1 *hold and let* $d_{y^\delta}(\lambda) := \mathcal{R}^*(A^*\lambda) - \langle\lambda, y^\delta\rangle$. *Let* $L := \|A\|^2/(2\sigma)$. *Consider the dual gradient method* (1.3) *with* $\lambda_0 = 0$. *If* $0 < \gamma \le 1/L$ *then for any* $\lambda \in Y$ *there holds*

$$
\begin{aligned}
d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{n+1}) &\ge \frac{1}{2\gamma(n+1)}\left(\|\lambda_{n+1} - \lambda\|^2 - \|\lambda\|^2\right) \\
&\quad + \left\{\left(\frac{1}{2} - \frac{L\gamma}{4}\right)n + \left(\frac{1}{2} - \frac{L\gamma}{2}\right)\right\}\gamma\|Ax_n - y^\delta\|^2
\end{aligned}
$$

*for all* $n \ge 0$.

**Proof** Since $\mathcal{R}$ is strongly convex with modulus of convexity $\sigma > 0$, it follows from Proposition 2.2 that $\mathcal{R}^*$ is continuously differentiable and

$$\|\nabla\mathcal{R}^*(\xi) - \nabla\mathcal{R}^*(\eta)\| \le \frac{\|\xi - \eta\|}{2\sigma}, \quad \forall\xi, \eta \in X^*.$$

Consequently, the function $\lambda \to d_{y^\delta}(\lambda)$ is differentiable on $Y$ and its gradient is given by

$$\nabla d_{y^\delta}(\lambda) = A \nabla \mathcal{R}^*(A^*\lambda) - y^\delta$$

with

$$\|\nabla d_{y^\delta}(\tilde{\lambda}) - \nabla d_{y^\delta}(\lambda)\| \le L\|\tilde{\lambda} - \lambda\|, \quad \forall \tilde{\lambda}, \lambda \in Y,$$

where $L = \|A\|^2/(2\sigma)$. Therefore

$$d_{y^\delta}(\lambda_{n+1}) \le d_{y^\delta}(\lambda_n) + \langle \nabla d_{y^\delta}(\lambda_n), \lambda_{n+1} - \lambda_n \rangle + \frac{L}{2}\|\lambda_{n+1} - \lambda_n\|^2.$$

By the convexity of $d_{y^\delta}$ we have for any $\lambda \in Y$ that

$$d_{y^\delta}(\lambda_n) \le d_{y^\delta}(\lambda) + \langle \nabla d_{y^\delta}(\lambda_n), \lambda_n - \lambda \rangle.$$

Combining the above equations we thus obtain

$$d_{y^\delta}(\lambda_{n+1}) \le d_{y^\delta}(\lambda) + \langle \nabla d_{y^\delta}(\lambda_n), \lambda_{n+1} - \lambda \rangle + \frac{L}{2}\|\lambda_{n+1} - \lambda_n\|^2.$$

By using (3.2) we can see $\nabla d_{y^\delta}(\lambda_n) = Ax_n - y^\delta$ which together with the equation $\lambda_{n+1} - \lambda_n = -\gamma(Ax_n - y^\delta)$ shows that $\nabla d_{y^\delta}(\lambda_n) = (\lambda_n - \lambda_{n+1})/\gamma$. Consequently

$$d_{y^\delta}(\lambda_{n+1}) \le d_{y^\delta}(\lambda) + \frac{1}{\gamma}\langle \lambda_n - \lambda_{n+1}, \lambda_{n+1} - \lambda \rangle + \frac{L}{2}\|\lambda_{n+1} - \lambda_n\|^2.$$

Note that

$$\langle \lambda_n - \lambda_{n+1}, \lambda_{n+1} - \lambda \rangle = \frac{1}{2}\left(\|\lambda_n - \lambda\|^2 - \|\lambda_{n+1} - \lambda\|^2 - \|\lambda_{n+1} - \lambda_n\|^2\right).$$

Therefore

$$d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{n+1}) \ge \frac{1}{2\gamma}\left(\|\lambda_{n+1} - \lambda\|^2 - \|\lambda_n - \lambda\|^2\right)$$

$$+ \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\|\lambda_{n+1} - \lambda_n\|^2. \tag{3.3}$$

Let $m \ge 0$ be any number. By summing (3.3) over $n$ from $n = 0$ to $n = m$ and using $\lambda_0 = 0$ we can obtain

$$\sum_{n=0}^{m} \left( d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{n+1}) \right) \geq \frac{1}{2\gamma} \left( \|\lambda_{m+1} - \lambda\|^2 - \|\lambda\|^2 \right)$$

$$+ \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \sum_{n=0}^{m} \|\lambda_{n+1} - \lambda_n\|^2. \qquad (3.4)$$

Next we take $\lambda = \lambda_n$ in (3.3) to obtain

$$d_{y^\delta}(\lambda_n) - d_{y^\delta}(\lambda_{n+1}) \geq \left( \frac{1}{\gamma} - \frac{L}{2} \right) \|\lambda_{n+1} - \lambda_n\|^2.$$

Multiplying this inequality by $n$ and then summing over $n$ from $n = 0$ to $n = m$ we can obtain

$$\sum_{n=0}^{m} n \left( d_{y^\delta}(\lambda_n) - d_{y^\delta}(\lambda_{n+1}) \right) \geq \left( \frac{1}{\gamma} - \frac{L}{2} \right) \sum_{n=0}^{m} n \|\lambda_{n+1} - \lambda_n\|^2.$$

Note that

$$\sum_{n=0}^{m} n \left( d_{y^\delta}(\lambda_n) - d_{y^\delta}(\lambda_{n+1}) \right) = -(m+1)d_{y^\delta}(\lambda_{m+1}) + \sum_{n=0}^{m} d_{y^\delta}(\lambda_{n+1}).$$

Thus

$$-(m+1)d_{y^\delta}(\lambda_{m+1}) + \sum_{n=0}^{m} d_{y^\delta}(\lambda_{n+1}) \geq \left( \frac{1}{\gamma} - \frac{L}{2} \right) \sum_{n=0}^{m} n \|\lambda_{n+1} - \lambda_n\|^2.$$

Adding this inequality to (3.4) gives

$$(m+1) \left( d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{m+1}) \right) \geq \frac{1}{2\gamma} \left( \|\lambda_{m+1} - \lambda\|^2 - \|\lambda\|^2 \right)$$

$$+ \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \sum_{n=0}^{m} \|\lambda_{n+1} - \lambda_n\|^2$$

$$+ \left( \frac{1}{\gamma} - \frac{L}{2} \right) \sum_{n=0}^{m} n \|\lambda_{n+1} - \lambda_n\|^2.$$

Recall that $\lambda_n - \lambda_{n+1} = \gamma(Ax_n - y^\delta)$. By using the monotonicity of $\|Ax_n - y^\delta\|$ shown in Lemma 3.2 we then obtain

$$(m+1) \left( d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{m+1}) \right)$$

$$\geq \frac{1}{2\gamma} \left( \|\lambda_{m+1} - \lambda\|^2 - \|\lambda\|^2 \right) + \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \gamma^2 \sum_{n=0}^{m} \|Ax_n - y^\delta\|^2$$

$$+ \left( \frac{1}{\gamma} - \frac{L}{2} \right) \gamma^2 \sum_{n=0}^{m} n \| A x_n - y^\delta \|^2$$

$$\geq \frac{1}{2\gamma} \left( \| \lambda_{m+1} - \lambda \|^2 - \| \lambda \|^2 \right) + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \gamma (m+1) \| A x_m - y^\delta \|^2$$

$$+ \left( 1 - \frac{L\gamma}{2} \right) \gamma \frac{m(m+1)}{2} \| A x_m - y^\delta \|^2.$$

The proof is therefore complete. □

We now assume that the unique solution $x^\dagger$ satisfies a variational source condition specified in the following assumption.

**Assumption 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

$$\mathcal{E}^\dagger(x) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + M \| A x - y \|^q, \quad \forall x \in \text{dom}(\mathcal{R})$$

for some $0 < q \leq 1$ and some constant $M > 0$.

Variational source conditions were first introduced in [23], 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, refined and verified, see [15, 17, 20, 24–26] for instance. The error measure function $\mathcal{E}^\dagger$ in Assumption 2 is used to measure the speed of convergence; it can be taken in various forms and the usual choice of $\mathcal{E}^\dagger$ is the Bregman distance induced by $\mathcal{R}$. Use of a general error measure functional has the advantage of covering a wider range of applications. For instance, in reconstructing sparse solutions of ill-posed problems, one may consider the sought solution in the $\ell^1$ space and take $\mathcal{R}(x) = \|x\|_{\ell^1}$. In this situation, convergence under the Bregman distance induced by $\mathcal{R}$ may not provide useful approximation result because two points with zero Bregman distance may have arbitrarily large $\ell^1$-distance. However, under certain natural conditions, the variational source conditions can be verified with $\mathcal{E}^\dagger(x) = \|x - x^\dagger\|_{\ell^1}$; see [15, 16].

We first derive the convergence rates for the dual gradient method (1.3) under an *a priori* stopping rule when $x^\dagger$ satisfies the variational source conditions specified in Assumption 2.

**Theorem 3.4** *Let Assumption 1 hold and let* $L := \|A\|^2/(2\sigma)$. *If* $0 < \gamma \leq 1/L$ *and* $x^\dagger$ *satisfies the variational source conditions specified in Assumption 2, then for the dual gradient method* (1.3) *with the initial guess* $\lambda_0 = 0$ *there holds*

$$\mathcal{E}^\dagger(x_n) \leq C \left( n^{-\frac{q}{2-q}} + \delta^q + n^{\frac{1-q}{2-q}} \delta + n \delta^2 \right)$$

*for all* $n \geq 1$, *where $C$ is a generic positive constant independent of $n$ and $\delta$. Consequently, by choosing an integer $n_\delta$ with $n_\delta \sim \delta^{q-2}$ we have*

$$\mathcal{E}^\dagger(x_{n_\delta}) = O(\delta^q).$$

**Proof** Let $d_y(\lambda) := \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle$. Since $0 < \gamma \le 1/L$, from Proposition 3.3 it follows that

$$\left\{ \left( \frac{1}{2} - \frac{L\gamma}{4} \right) n + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \right\} \gamma \|Ax_n - y^\delta\|^2$$

$$\le d_{y^\delta}(\lambda) - d_{y^\delta}(\lambda_{n+1}) - \frac{1}{2\gamma(n+1)} \left( \|\lambda_{n+1} - \lambda\|^2 - \|\lambda\|^2 \right)$$

$$= d_y(\lambda) - d_y(\lambda_{n+1}) + \langle \lambda_{n+1} - \lambda, y^\delta - y \rangle$$

$$- \frac{1}{2\gamma(n+1)} \left( \|\lambda_{n+1} - \lambda\|^2 - \|\lambda\|^2 \right) \tag{3.5}$$

for all $\lambda \in Y$. By the Cauchy-Schwarz inequality we have

$$\langle \lambda_{n+1} - \lambda, y^\delta - y \rangle \le \delta \|\lambda_{n+1} - \lambda\| \le \frac{1}{4\gamma(n+1)} \|\lambda_{n+1} - \lambda\|^2 + \gamma(n+1)\delta^2.$$

Thus, it follows from (3.5) that

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{4\gamma(n+1)} \|\lambda_{n+1} - \lambda\|^2$$

$$\le d_y(\lambda) - d_y(\lambda_{n+1}) + \frac{\|\lambda\|^2}{2\gamma(n+1)} + \gamma(n+1)\delta^2,$$

where $c_0 := (1/2 - L\gamma/4)\gamma > 0$. By virtue of the inequality $\|\lambda_{n+1}\|^2 \le 2(\|\lambda\|^2 + \|\lambda_{n+1} - \lambda\|^2)$ we then have

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{8\gamma(n+1)} \|\lambda_{n+1}\|^2$$

$$\le d_y(\lambda) - d_y(\lambda_{n+1}) + \frac{3\|\lambda\|^2}{4\gamma(n+1)} + \gamma(n+1)\delta^2.$$

By the Fenchel-Young inequality (2.1) and $Ax^\dagger = y$ we have

$$d_y(\lambda_{n+1}) = \mathcal{R}^*(A^*\lambda_{n+1}) - \langle \lambda_{n+1}, Ax^\dagger \rangle$$

$$= \mathcal{R}^*(A^*\lambda_{n+1}) - \langle A^*\lambda_{n+1}, x^\dagger \rangle$$

$$\ge -\mathcal{R}(x^\dagger). \tag{3.6}$$

Therefore

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{8\gamma(n+1)} \|\lambda_{n+1}\|^2$$

$$\le \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^\dagger) + \frac{3\|\lambda\|^2}{4\gamma(n+1)} + \gamma(n+1)\delta^2$$

for all $\lambda \in Y$. Consequently

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{8\gamma(n+1)} \|\lambda_{n+1}\|^2$$

$$\leq \inf_{\lambda \in Y} \left\{ \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^\dagger) + \frac{3\|\lambda\|^2}{4\gamma(n+1)} \right\} + \gamma(n+1)\delta^2$$

$$= \mathcal{R}(x^\dagger) - \sup_{\lambda \in Y} \left\{ -\mathcal{R}^*(A^*\lambda) + \langle \lambda, y \rangle - \frac{3\|\lambda\|^2}{4\gamma(n+1)} \right\} + \gamma(n+1)\delta^2.$$

According to the Fenchel–Rockafellar duality formula given in Proposition 2.1, we have

$$\sup_{\lambda \in Y} \left\{ -\mathcal{R}^*(A^*\lambda) + \langle \lambda, y \rangle - \frac{3\|\lambda\|^2}{4\gamma(n+1)} \right\} = \inf_{x \in X} \left\{ \mathcal{R}(x) + \frac{1}{3}\gamma(n+1)\|Ax - y\|^2 \right\}.$$

Indeed, by taking $f(x) = \mathcal{R}(x)$ for $x \in X$ and $g(z) = \frac{1}{3}\gamma(n+1)\|z - y\|^2$ for $z \in Y$, we can obtain this identity immediately from (2.3) by noting that

$$g^*(\lambda) = \frac{3}{4\gamma(n+1)} \|\lambda\|^2 + \langle \lambda, y \rangle, \quad \lambda \in Y.$$

Therefore

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{8\gamma(n+1)} \|\lambda_{n+1}\|^2 \leq \eta_n + \gamma(n+1)\delta^2, \qquad (3.7)$$

where

$$\eta_n := \sup_{x \in X} \left\{ \mathcal{R}(x^\dagger) - \mathcal{R}(x) - \frac{1}{3}\gamma(n+1)\|Ax - y\|^2 \right\}. \qquad (3.8)$$

We now estimate $\eta_n$ when $x^\dagger$ satisfies the variatioinal source condition given in Assumption 2. By the nonnegativity of $\mathcal{E}^\dagger$ we have $\mathcal{R}(x^\dagger) - \mathcal{R}(x) \leq M\|Ax - y\|^q$. Thus

$$\eta_n \leq \sup_{x \in X} \left\{ M\|Ax - y\|^q - \frac{1}{3}\gamma(n+1)\|Ax - y\|^2 \right\}$$

$$\leq \sup_{s \geq 0} \left\{ Ms^q - \frac{1}{3}\gamma(n+1)s^2 \right\} = c_1(n+1)^{-\frac{q}{2-q}}, \qquad (3.9)$$

where $c_1 := \left(1 - \frac{q}{2}\right)\left(\frac{3qM}{2\gamma}\right)^{\frac{q}{2-q}} M > 0$. Combining this with (3.7) gives

$$c_0 n \|Ax_n - y^\delta\|^2 + \frac{1}{8\gamma(n+1)} \|\lambda_{n+1}\|^2 \leq c_1(n+1)^{-\frac{q}{2-q}} + \gamma(n+1)\delta^2$$

which implies that

$$\|Ax_n - y^\delta\| \le C\left(n^{-\frac{1}{2-q}} + \delta\right) \quad \text{and} \quad \|\lambda_n\| \le C\left(n^{\frac{1-q}{2-q}} + n\delta\right). \tag{3.10}$$

Recall $A^*\lambda_n \in \partial\mathcal{R}(x_n)$ from (3.1), we have

$$\mathcal{R}(x_n) - \mathcal{R}(x^\dagger) \le \langle A^*\lambda_n, x_n - x^\dagger \rangle = \langle \lambda_n, Ax_n - y \rangle.$$

Therefore, by using the variational source condition specified in Assumption 2, we obtain

$$\begin{aligned}
\mathcal{E}^\dagger(x_n) &\le \mathcal{R}(x_n) - \mathcal{R}(x^\dagger) + M\|Ax_n - y\|^q \\
&\le \langle \lambda_n, Ax_n - y \rangle + M\|Ax_n - y\|^q \\
&\le \|\lambda_n\|\|Ax_n - y\| + M\|Ax_n - y\|^q.
\end{aligned}$$

Thus, it follows from (3.10) that

$$\begin{aligned}
\mathcal{E}^\dagger(x_n) &\le \|\lambda_n\| \left(\|Ax_n - y^\delta\| + \delta\right) + M\left(\|Ax_n - y^\delta\| + \delta\right)^q \\
&\le C\left(n^{\frac{1-q}{2-q}} + n\delta\right)\left(n^{-\frac{1}{2-q}} + \delta\right) + C\left(n^{-\frac{1}{2-q}} + \delta\right)^q \\
&\le C\left(n^{-\frac{q}{2-q}} + \delta^q + n^{\frac{1-q}{2-q}}\delta + n\delta^2\right).
\end{aligned}$$

The proof is thus complete.                                                                              $\square$

During the Proof of Theorem 3.4, we have introduced the quantity $\eta_n$ defined by (3.8). Taking $x = x^\dagger$ in (3.8) shows $\eta_n \ge 0$. As can be seen from the proof of Theorem 3.4, we have

$$\eta_n = \inf_{\lambda \in Y} \left\{ \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^\dagger) + \frac{3\|\lambda\|^2}{4\gamma(n+1)} \right\}$$

by the Fenchel-Rockafellar duality formula. Taking $\lambda = 0$ in this equation gives $0 \le \eta_n \le \mathcal{R}(x^\dagger) + \mathcal{R}^*(0) < \infty$. The proof of Theorem 3.4 demonstrates that $\eta_n$ can decay to 0 at certain rate if $x^\dagger$ satisfies a variational source condition.

**Corollary 3.5** *Let Assumption 1 hold and let $L := \|A\|^2/(2\sigma)$. If $0 < \gamma \le 1/L$ and if there is $\lambda^\dagger \in Y$ such that $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$, then for the dual gradient method (1.3) with the initial guess $\lambda_0 = 0$ there holds*

$$D_{\mathcal{R}}^{A^*\lambda^\dagger}(x_n, x^\dagger) \le C\left(n^{-1} + \delta + n\delta^2\right) \tag{3.11}$$

for all $n \geq 1$, where $C$ is a generic positive constant independent of $n$ and $\delta$. Consequently, by choosing an integer $n_\delta$ with $n_\delta \sim \delta^{-1}$ we have

$$D_{\mathcal{R}}^{A^*\lambda^\dagger}(x_{n_\delta}, x^\dagger) = O(\delta) \tag{3.12}$$

and hence $\|x_{n_\delta} - x^\dagger\| = O(\delta^{1/2})$.

**Proof** We show that $x^\dagger$ satisfies the variational source condition specified in Assumption 2 with $q = 1$. The argument is well-known, see [23] for instance. Since $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$ for some $\lambda^\dagger \in Y$, we have for all $x \in \text{dom}(\mathcal{R})$ that

$$D_{\mathcal{R}}^{A^*\lambda^\dagger}(x, x^\dagger) = \mathcal{R}(x) - \mathcal{R}(x^\dagger) - \langle \lambda^\dagger, Ax - y \rangle$$
$$\leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + \|\lambda^\dagger\|\|Ax - y\|$$

which shows that Assumption 2 holds with $\mathcal{E}^\dagger(x) = D_{\mathcal{R}}^{A^*\lambda^\dagger}(x, x^\dagger)$, $M = \|\lambda^\dagger\|$ and $q = 1$. Thus by invoking Theorem 3.4, we immediately obtain (3.11) which together with the choice $n_\delta \sim \delta^{-1}$ implies (3.12). By using (2.5) we then obtain $\|x_{n_\delta} - x^\dagger\| = O(\delta^{1/2})$. $\qquad\square$

We next turn to deriving convergence rates of the dual gradient method (1.3) under the variational source condition given in Assumption 2 when the method is terminated by the discrepancy principle (1.5). We will use the following consequence of Proposition 3.3.

**Lemma 3.6** Let Assumption 1 hold and let $L := \|A\|^2/(2\sigma)$. Consider the dual gradient method (1.3) with $\lambda_0 = 0$. If $\tau > 1$ and $\gamma > 0$ are chosen such that $1 - 1/\tau^2 - L\gamma > 0$, then there is a constant $c_2 > 0$ such that

$$c_2(n+1)\delta^2 \leq \eta_n \quad and \quad \frac{1}{8\gamma(n+1)}\|\lambda_{n+1}\|^2 \leq \eta_n + \gamma(n+1)\delta^2$$

for all integers $0 \leq n < n_\delta$, where $n_\delta$ is the integer determined by the discrepancy principle (1.5) and $\eta_n$ is the quantity defined by (3.8).

**Proof** The second estimate follows directly from (3.7), actually it holds for all integers $n \geq 0$. It remains only to show the first estimate. For any $n < n_\delta$ we have $\|Ax_n - y^\delta\| > \tau\delta$. Therefore from (3.5) it follows for all $\lambda \in Y$ that

$$\left\{ \left( \frac{1}{2} - \frac{L\gamma}{4} \right) n + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \right\} \gamma\tau^2\delta^2$$
$$\leq d_y(\lambda) - d_y(\lambda_{n+1}) + \langle \lambda_{n+1} - \lambda, y^\delta - y \rangle - \frac{1}{2\gamma(n+1)} \left( \|\lambda_{n+1} - \lambda\|^2 - \|\lambda\|^2 \right).$$

By the Cauchy-Schwarz inequality we have

$$\langle \lambda_{n+1} - \lambda, y^\delta - y \rangle$$
$$\leq \delta \|\lambda_{n+1} - \lambda\| \leq \frac{1}{2\gamma(n+1)} \|\lambda_{n+1} - \lambda\|^2 + \frac{1}{2}\gamma(n+1)\delta^2.$$

Therefore

$$\left[ \left\{ \left( \frac{1}{2} - \frac{L\gamma}{4} \right) n + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \right\} \tau^2 - \frac{1}{2}(n+1) \right] \gamma \delta^2$$
$$\leq d_y(\lambda) - d_y(\lambda_{n+1}) + \frac{1}{2\gamma(n+1)} \|\lambda\|^2.$$

By the conditions on $\gamma$ and $\tau$, it is easy to see that

$$\left[ \left\{ \left( \frac{1}{2} - \frac{L\gamma}{4} \right) n + \left( \frac{1}{2} - \frac{L\gamma}{2} \right) \right\} \tau^2 - \frac{1}{2}(n+1) \right] \gamma \geq c_2(n+1),$$

where $c_2 := \left( (1/2 - L\gamma/2)\tau^2 - 1/2 \right) \gamma > 0$. Therefore

$$c_2(n+1)\delta^2 \leq d_y(\lambda) - d_y(\lambda_{n+1}) + \frac{1}{2\gamma(n+1)} \|\lambda\|^2.$$

According to (3.6) we have $d_y(\lambda_{n+1}) \geq -\mathcal{R}(x^\dagger)$. Thus

$$c_2(n+1)\delta^2 \leq \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^\dagger) + \frac{1}{2\gamma(n+1)} \|\lambda\|^2$$

which is valid for all $\lambda \in Y$. Consequently

$$c_2(n+1)\delta^2 \leq \inf_{\lambda \in Y} \left\{ \mathcal{R}^*(A^*\lambda) - \langle \lambda, y \rangle + \mathcal{R}(x^\dagger) + \frac{1}{2\gamma(n+1)} \|\lambda\|^2 \right\}$$
$$= \mathcal{R}(x^\dagger) - \sup_{\lambda \in Y} \left\{ -\mathcal{R}^*(A^*\lambda) + \langle \lambda, y \rangle - \frac{1}{2\gamma(n+1)} \|\lambda\|^2 \right\}.$$

According to the Fenchel-Rockafellar duality formula given in Proposition 2.1, we can further obtain

$$c_2(n+1)\delta^2 \leq \mathcal{R}(x^\dagger) - \inf_{x \in X} \left\{ \mathcal{R}(x) + \frac{1}{2}\gamma(n+1)\|Ax - y\|^2 \right\}$$
$$= \sup_{x \in X} \left\{ \mathcal{R}(x^\dagger) - \mathcal{R}(x) - \frac{1}{2}\gamma(n+1)\|Ax - y\|^2 \right\}$$
$$\leq \eta_n$$

which shows the first estimate. $\qquad\square$

Now we are ready to show the convergence rate result for the dual gradient method (1.3) under Assumption 2 when the method is terminated by the discrepancy principle (1.5).

**Theorem 3.7** *Let Assumption* 1 *hold and let* $L := \|A\|^2/(2\sigma)$. *Consider the dual gradient method* (1.3) *with the initial guess* $\lambda_0 = 0$. *Assume that* $\tau > 1$ *and* $\gamma > 0$ *are chosen such that* $1 - 1/\tau^2 - L\gamma > 0$ *and let* $n_\delta$ *be the integer determined by the discrepancy principle* (1.5). *If* $x^\dagger$ *satisfies the variational source condition specified in Assumption* 2, *then*

$$\mathcal{E}^\dagger(x_{n_\delta}) = O(\delta^q). \tag{3.13}$$

*Consequently, if there is* $\lambda^\dagger \in Y$ *such that* $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$, *then*

$$D_\mathcal{R}^{A^*\lambda^\dagger}(x_{n_\delta}, x^\dagger) = O(\delta) \tag{3.14}$$

*and hence* $\|x_{n_\delta} - x^\dagger\| = O(\delta^{1/2})$.

**Proof** By using the variational source condition on $x^\dagger$ specified in Assumption 2, the convexity of $\mathcal{R}$, and the fact $A^*\lambda_{n_\delta} \in \partial\mathcal{R}(x_{n_\delta})$ we have

$$\begin{aligned}
\mathcal{E}^\dagger(x_{n_\delta}) &\le \mathcal{R}(x_{n_\delta}) - \mathcal{R}(x^\dagger) + M\|Ax_{n_\delta} - y\|^q \\
&\le \langle \lambda_{n_\delta}, Ax_{n_\delta} - y \rangle + M\|Ax_{n_\delta} - y\|^q \\
&\le \|\lambda_{n_\delta}\|\|Ax_{n_\delta} - y\| + M\|Ax_{n_\delta} - y\|^q.
\end{aligned}$$

By the definition of $n_\delta$ we have $\|Ax_{n_\delta} - y^\delta\| \le \tau\delta$ and thus

$$\|Ax_{n_\delta} - y\| \le \|Ax_{n_\delta} - y^\delta\| + \|y^\delta - y\| \le (\tau + 1)\delta.$$

Therefore

$$\mathcal{E}^\dagger(x_{n_\delta}) \le (\tau + 1)\|\lambda_{n_\delta}\|\delta + M(\tau + 1)^q\delta^q. \tag{3.15}$$

If $n_\delta = 0$, then we have $\lambda_{n_\delta} = 0$ and hence $\mathcal{E}^\dagger(x_{n_\delta}) \le M(\tau + 1)^q\delta^q$. In the following we consider the case $n_\delta \ge 1$. We will use Lemma 3.6 to estimate $\|\lambda_{n_\delta}\|$. By virtue of Assumption 2 we have $\eta_n \le c_1(n + 1)^{-\frac{q}{2-q}}$, see (3.9). Combining this with the estimates in Lemma 3.6 we can obtain

$$c_2(n + 1)^{\frac{2}{2-q}}\delta^2 \le c_1 \tag{3.16}$$

and

$$\|\lambda_{n+1}\|^2 \le 8\gamma c_1(n + 1)^{\frac{2(1-q)}{2-q}} + 8\gamma^2(n + 1)^2\delta^2 \tag{3.17}$$

for all $0 \le n < n_\delta$. Taking $n = n_\delta - 1$ in (3.16) gives

$$n_\delta \le \left( \frac{c_1}{c_2 \delta^2} \right)^{\frac{2-q}{2}}$$

which together with (3.17) with $n = n_\delta - 1$ shows that

$$\|\lambda_{n_\delta}\| \le c_3 \delta^{q-1},$$

where $c_3 := \sqrt{8\gamma(\gamma + c_2)(c_1/c_2)^{2-q}}$. Combining this estimate with (3.15) we finally obtain

$$\mathcal{E}^\dagger(x_{n_\delta}) \le \left( c_3(\tau + 1) + M(\tau + 1)^q \right) \delta^q$$

which shows (3.13).

When $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$ for some $\lambda^\dagger \in Y$, we know from the proof of Corollary 3.5 that Assumption 2 is satisfied with $\mathcal{E}^\dagger(x) = D_\mathcal{R}^{A^*\lambda^\dagger}(x, x^\dagger)$ and $q = 1$. Thus, we may use (3.13) to conclude (3.14). □

### 3.3 Acceleration

The dual gradient method, which generalizes the linear Landweber iteration in Hilbert spaces, is a slowly convergent method in general. To make it more practical important, it is necessary to consider accelerating this method with faster convergence speed. Since the dual gradient method is obtained by applying the gradient method to the dual problem, one may consider to accelerate this method by applying any available acceleration strategy for gradient methods among which Nesterov's acceleration strategy [2, 6, 37] is the most prominent. By applying Nesterov's accelerated gradient method to minimize the function $d_{y^\delta}(\lambda) = \mathcal{R}^*(A^*\lambda) - \langle \lambda, y^\delta \rangle$ it leads to the iteration scheme

$$\hat{\lambda}_n = \lambda_n + \frac{n-1}{n+\alpha}(\lambda_n - \lambda_{n-1}),$$
$$\lambda_{n+1} = \hat{\lambda}_n - \gamma \nabla d_{y^\delta}(\hat{\lambda}_n).$$

Let $\hat{x}_n = \nabla\mathcal{R}^*(A^*\hat{\lambda}_n)$. Then $\nabla d_{y^\delta}(\hat{\lambda}_n) = A\hat{x}_n - y^\delta$ and $A^*\hat{\lambda}_n \in \partial\mathcal{R}(\hat{x}_n)$ which imply that

$$\hat{\lambda}_n = \lambda_n + \frac{n-1}{n+\alpha}(\lambda_n - \lambda_{n-1}),$$
$$\hat{x}_n = \arg\min_{x \in X}\left\{ \mathcal{R}(x) - \langle \hat{\lambda}_n, Ax - y^\delta \rangle \right\},$$
$$\lambda_{n+1} = \hat{\lambda}_n - \gamma(A\hat{x}_n - y^\delta),$$
$$x_{n+1} = \arg\min_{x \in X}\left\{ \mathcal{R}(x) - \langle \lambda_{n+1}, Ax - y^\delta \rangle \right\}, \tag{3.18}$$

where $\lambda_{-1} = \lambda_0 = 0$, $\alpha \geq 2$ is a given number, and $\gamma > 0$ is a step size. We have the following convergence rate result when the method is terminated by an *a priori* stopping rule.

**Theorem 3.8** *Let Assumption 1 hold and let $L := \|A\|^2/(2\sigma)$. Consider the accelerated dual gradient method (3.18) with noisy data $y^\delta$ satisfying $\|y^\delta - y\| \leq \delta$. Assume that $0 < \gamma \leq 1/L$ and $\alpha \geq 2$. If $x^\dagger$ satisfies the source condition $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$ for some $\lambda^\dagger \in Y$, then there exist positive constants $c_4$ and $c_5$ depending only on $\gamma$ and $\alpha$ such that*

$$D_\mathcal{R}^{A^*\lambda_n}(x^\dagger, x_n) \leq \left(\frac{c_4\|\lambda^\dagger\|}{n} + c_5 n\delta\right)^2 \tag{3.19}$$

*for all $n \geq 1$. Consequently by choosing an integer $n_\delta$ with $n_\delta \sim \delta^{-1/2}$ we have*

$$D_\mathcal{R}^{A^*\lambda_{n_\delta}}(x^\dagger, x_{n_\delta}) = O(\delta)$$

*and hence $\|x_{n_\delta} - x^\dagger\| = O(\delta^{1/2})$ as $\delta \to 0$.*

**Proof** According to the definition of $x_n$ we have $A^*\lambda_n \in \partial\mathcal{R}(x_n)$ for all $n \geq 1$. From this fact and the condition $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$ it follows from (2.2) that

$$
\begin{aligned}
D_\mathcal{R}^{A^*\lambda_n}(x^\dagger, x_n) &= \mathcal{R}(x^\dagger) - \mathcal{R}(x_n) - \langle\lambda_n, y - Ax_n\rangle \\
&= \left\{\langle A^*\lambda^\dagger, x^\dagger\rangle - \mathcal{R}^*(A^*\lambda^\dagger)\right\} - \left\{\langle A^*\lambda_n, x_n\rangle - \mathcal{R}^*(A^*\lambda_n)\right\} \\
&\quad - \langle\lambda_n, y - Ax_n\rangle \\
&= \mathcal{R}^*(A^*\lambda_n) - \mathcal{R}^*(A^*\lambda^\dagger) - \langle\lambda_n, y\rangle + \langle\lambda^\dagger, y\rangle \\
&= d_y(\lambda_n) - d_y(\lambda^\dagger),
\end{aligned}
\tag{3.20}
$$

where $d_y(\lambda) := \mathcal{R}^*(A^*\lambda) - \langle\lambda, y\rangle$. We need to estimate $d_y(\lambda_n) - d_y(\lambda^\dagger)$. This can be done by using a perturbation analysis of the accelerated gradient method, see [2, 3]. For completeness, we include a derivation here. Because $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$, we have $x^\dagger = \nabla\mathcal{R}^*(A^*\lambda^\dagger)$. Thus

$$\nabla d_y(\lambda^\dagger) = A\nabla\mathcal{R}^*(A^*\lambda^\dagger) - y = Ax^\dagger - y = 0.$$

Since $d_y$ is convex, this shows that $\lambda^\dagger$ is a global minimizer of $d_y$ over $Y$. Note that $\nabla d_{y^\delta}(\lambda) = \nabla d_y(\lambda) + y - y^\delta$. Thus, it follows from the definition of $\lambda_{n+1}$ that

$$\lambda_{n+1} = \hat\lambda_n - \gamma\left(\nabla d_y(\hat\lambda_n) + y - y^\delta\right).$$

Based on this, the Lipschitz continuity of $\nabla d_y$ and the convexity of $\mathcal{R}$, we may use a similar argument in the proof of Proposition 3.3 to obtain for any $\lambda \in Y$ that

$$d_y(\lambda_{n+1}) \leq d_y(\lambda) + \langle \nabla d_y(\hat{\lambda}_n), \lambda_{n+1} - \lambda \rangle + \frac{L}{2} \|\lambda_{n+1} - \hat{\lambda}_n\|^2$$

$$= d_y(\lambda) + \left\langle \frac{1}{\gamma}(\hat{\lambda}_n - \lambda_{n+1}) - (y - y^\delta), \lambda_{n+1} - \lambda \right\rangle + \frac{L}{2} \|\lambda_{n+1} - \hat{\lambda}_n\|^2$$

$$= d_y(\lambda) + \frac{1}{2\gamma} \left( \|\hat{\lambda}_n - \lambda\|^2 - \|\lambda_{n+1} - \lambda\|^2 \right) - \langle y - y^\delta, \lambda_{n+1} - \lambda \rangle$$

$$- \left( \frac{1}{2\gamma} - \frac{L}{2} \right) \|\lambda_{n+1} - \hat{\lambda}_n\|^2.$$

Since $0 < \gamma \leq 1/L$ and $\|y^\delta - y\| \leq \delta$, we have

$$d_y(\lambda_{n+1}) \leq d_y(\lambda) + \frac{1}{2\gamma} \left( \|\hat{\lambda}_n - \lambda\|^2 - \|\lambda_{n+1} - \lambda\|^2 \right) + \delta \|\lambda_{n+1} - \lambda\|. \quad (3.21)$$

Note that $\frac{n-1}{n+\alpha} = \frac{t_n - 1}{t_{n+1}}$ with $t_n = \frac{n+\alpha-1}{\alpha}$. Now we take $\lambda = \left( 1 - \frac{1}{t_{n+1}} \right) \lambda_n + \frac{1}{t_{n+1}} \lambda^\dagger$ in (3.21) and use the convexity of $d_y$ to obtain

$$d_y(\lambda_{n+1}) \leq \left( 1 - \frac{1}{t_{n+1}} \right) d_y(\lambda_n) + \frac{1}{t_{n+1}} d_y(\lambda^\dagger)$$

$$+ \frac{1}{2\gamma t_{n+1}^2} \left\| \lambda^\dagger - \left( \lambda_n + t_{n+1}(\hat{\lambda}_n - \lambda_n) \right) \right\|^2$$

$$- \frac{1}{2\gamma t_{n+1}^2} \left\| \lambda^\dagger - (\lambda_n + t_{n+1}(\lambda_{n+1} - \lambda_n)) \right\|^2$$

$$+ \frac{\delta}{t_{n+1}} \left\| \lambda^\dagger - (\lambda_n + t_{n+1}(\lambda_{n+1} - \lambda_n)) \right\|.$$

Let $u_n = \lambda_{n-1} + t_n(\lambda_n - \lambda_{n-1})$. Then it follows from $\hat{\lambda}_n = \lambda_n + \frac{t_n - 1}{t_{n+1}}(\lambda_n - \lambda_{n-1})$ that $\lambda_n + t_{n+1}(\hat{\lambda}_n - \lambda_n) = u_n$. Therefore

$$d_y(\lambda_{n+1}) \leq \left( 1 - \frac{1}{t_{n+1}} \right) d_y(\lambda_n) + \frac{1}{t_{n+1}} d_y(\lambda^\dagger) + \frac{1}{2\gamma t_{n+1}^2} \|\lambda^\dagger - u_n\|^2$$

$$- \frac{1}{2\gamma t_{n+1}^2} \|\lambda^\dagger - u_{n+1}\|^2 + \frac{\delta}{t_{n+1}} \|\lambda^\dagger - u_{n+1}\|.$$

Multiplying both sides by $2\gamma t_{n+1}^2$, regrouping the terms and setting $w_n := d_y(\lambda_n) - d_y(\lambda^\dagger)$, we obtain

$$2\gamma t_{n+1}^2 w_{n+1} - 2\gamma t_n^2 w_n \leq 2\gamma \rho_n w_n + \|\lambda^\dagger - u_n\|^2 - \|\lambda^\dagger - u_{n+1}\|^2$$

$$+ 2\gamma \delta t_{n+1} \|\lambda^\dagger - u_{n+1}\| \quad (3.22)$$

for all $n \geq 0$, where $\rho_n := t_{n+1}^2 - t_{n+1} - t_n^2$. Note that $\alpha \geq 2$ implies $\rho_n \leq 0$ for $n \geq 1$. Let $m \geq 1$ be any integer. Summing the above inequality over $n$ from $n = 1$ to $n = m - 1$ and using $w_n \geq 0$, we can obtain

$$2\gamma t_m^2 w_m + \|\lambda^\dagger - u_m\|^2 \leq 2\gamma t_1^2 w_1 + \|\lambda^\dagger - u_1\|^2 + 2\gamma\delta \sum_{k=2}^{m} t_k \|\lambda^\dagger - u_k\|.$$

Using (3.22) with $n = 0$ and noting that $t_0^2 + \rho_0 = 0$ and $u_0 = 0$ we can further obtain

$$2\gamma t_m^2 w_m + \|\lambda^\dagger - u_m\|^2$$
$$\leq 2\gamma(t_0^2 + \rho_0)w_0 + \|\lambda^\dagger\|^2 + 2\gamma\delta t_1 \|\lambda^\dagger - u_1\| + 2\gamma\delta \sum_{k=2}^{m} t_k \|\lambda^\dagger - u_k\|$$
$$= \|\lambda^\dagger\|^2 + 2\gamma\delta \sum_{k=1}^{m} t_k \|\lambda^\dagger - u_k\|. \tag{3.23}$$

According to (3.23), we have

$$\|\lambda^\dagger - u_m\|^2 \leq \|\lambda^\dagger\|^2 + 2\gamma\delta \sum_{k=1}^{m} t_k \|\lambda^\dagger - u_k\| \tag{3.24}$$

from which we may use an induction argument to obtain

$$\|\lambda^\dagger - u_m\| \leq \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{m} t_k \tag{3.25}$$

for all integers $m \geq 0$. Indeed, since $u_0 = 0$, (3.25) holds trivially for $m = 0$. Assume next that (3.25) holds for all $0 \leq m \leq n$ for some $n \geq 0$. We show (3.25) also holds for $m = n + 1$. If there is $0 \leq m \leq n$ such that $\|\lambda^\dagger - u_{n+1}\| \leq \|\lambda^\dagger - u_m\|$, then by the induction hypothesis we have

$$\|\lambda^\dagger - u_{n+1}\| \leq \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{m} t_k \leq \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{n+1} t_k.$$

So we may assume $\|\lambda^\dagger - u_{n+1}\| > \|\lambda^\dagger - u_m\|$ for all $0 \leq m \leq n$. It then follows from (3.24) that

$$\|\lambda^\dagger - u_{n+1}\|^2 \leq \|\lambda^\dagger\|^2 + 2\gamma\delta \left(\sum_{k=1}^{n+1} t_k\right) \|\lambda^\dagger - u_{n+1}\|.$$

By using the elementary inequality "$a^2 \leq b^2 + ca \implies a \leq b + c$ for $a, b, c \geq 0$", we obtain again

$$\|\lambda^\dagger - u_{n+1}\| \leq \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{n+1} t_k.$$

By the induction principle, we thus obtain (3.25). Based on (3.23) and (3.25) we have

$$
\begin{aligned}
2\gamma t_m^2 w_m &\leq \|\lambda^\dagger\|^2 + 2\gamma\delta \sum_{k=1}^{m} t_k \|\lambda^\dagger - u_k\| \\
&\leq \|\lambda^\dagger\|^2 + 2\gamma\delta \left( \sum_{k=1}^{m} t_k \right) \left( \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{m} t_k \right) \\
&\leq \left( \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{m} t_k \right)^2 .
\end{aligned}
$$

Thus, by the definition of $t_n$ it is straightforward to see that

$$d_y(\lambda_m) - d_y(\lambda^\dagger) \leq \frac{1}{2\gamma t_m^2} \left( \|\lambda^\dagger\| + 2\gamma\delta \sum_{k=1}^{m} t_k \right)^2 \leq \left( \frac{c_4 \|\lambda^\dagger\|}{m} + c_5 m\delta \right)^2,$$

where $c_4$ and $c_5$ are two positive constants depending only on $\gamma$ and $\alpha$. Combining this with (3.20) we thus complete the proof of (3.19).                                  □

From Theorem 3.8 it follows that, under the source condition $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$, we can obtain the convergence rate $\|x_{n_\delta} - x^\dagger\| = O(\delta^{1/2})$ within $O(\delta^{-1/2})$ iterations for the method (3.18). For the dual gradient method (1.3), however, we need to perform $O(\delta^{-1})$ iterations to achieve the same convergence rate, see Corollary 3.5. This demonstrates that the method (3.18) indeed has acceleration effect.

We remark that Nesterov's acceleration strategy was first proposed in [29] to accelerate gradient type regularization method for linear as well as nonlinear ill-posed problems in Banach spaces and various numerical results were reported which demonstrate the striking performance; see also [27, 28, 35, 39, 45] for further numerical simulations. Although we have proved in Theorem 3.8 a convergence rate result for the method (3.18) under an *a priori* stopping rule, the regularization property of the method under the discrepancy principle is not yet established for general strongly convex $\mathcal{R}$. However, when $X$ is a Hilbert space and $\mathcal{R}(x) = \|x\|^2/2$, the regularization property of the corresponding method has been established in [35, 39] based on a general acceleration framework in [21] using orthogonal polynomials; in particular it was observed in [35] that the parameter $\alpha$ plays an interesting role in deriving order optimal convergence rates. For an analysis of Nesterov's acceleration for nonlinear ill-posed problems in Hilbert spaces, one may refer to [28].

## 4 Applications

Various applications of the dual gradient method (1.3), or equivalently the method (1.4), have been considered in [9, 29, 33] for sparsity recovery and image reconstruction through the choices of $\mathcal{R}$ as strong convex perturbations of the $L^1$ and the total variation functionals and the numerical results demonstrate its nice performance. In the following we will provide some additional applications.

### 4.1 Dual projected landweber iteration

We first consider the application of our convergence theory to linear ill-posed problems in Hilbert spaces with convex constraint. Such problems arise from a number of real applications including the computed tomography [36] in which the sought solutions are nonnegative.

Let $A : X \to Y$ be a bounded linear operator between two Hilbert spaces $X$ and $Y$ and let $C \subset X$ be a closed convex set. Given $y \in Y$ and assuming that $Ax = y$ has a solution in $C$, we consider finding the unique solution $x^\dagger$ of $Ax = y$ in $C$ with minimal norm which can be stated as the minimization problem

$$\min \left\{ \frac{1}{2} \|x\|^2 : x \in C \text{ and } Ax = y \right\}. \tag{4.1}$$

This problem takes the form (1.1) with $\mathcal{R}(x) := \frac{1}{2}\|x\|^2 + \delta_C(x)$, where $\delta_C$ denotes the indicator function of $C$, i.e. $\delta_C(x) = 0$ if $x \in C$ and $\infty$ otherwise. Clearly $\mathcal{R}$ satisfies Assumption 1 (ii). It is easy to see that for any $\xi \in X$ the unique solution of

$$\min_{x \in X} \{\mathcal{R}(x) - \langle \xi, x \rangle\}$$

is given by $P_C(\xi)$, where $P_C$ denotes the metric projection of $X$ onto $C$. Therefore, applying the algorithm (1.3) to (4.1) leads to the dual projected Landweber iteration

$$\begin{aligned} x_n &= P_C(A^*\lambda_n), \\ \lambda_{n+1} &= \lambda_n - \gamma(Ax_n - y^\delta) \end{aligned} \tag{4.2}$$

that has been considered in [12]. Besides a stability estimate, it has been shown in [12] that, for the method (4.2) with exact data $y^\delta = y$, if $x^\dagger \in P_C(A^*Y)$ then

$$\sum_{n=1}^{\infty} \|x_n - x^\dagger\|^2 < \infty$$

which implies $\|x_n - x^\dagger\| \to 0$ as $n \to \infty$ but does not provide an error estimate unless $\|x_n - x^\dagger\|$ is monotonically decreasing which is unfortunately unknown. Therefore, the work in [12] does not tell much information about the regularization property of the method (4.2). It is natural to ask if the method (4.2) renders a regularization

method under *a priori* or *a posteriori* stopping rules and if it is possible to derive error estimates under suitable source conditions on the sought solution. Applying our convergence theory can provide satisfactory answers to these questions with a rather complete analysis on the method (4.2), see Corollary 4.1 below, which is far beyond the one provided in [12]. In particular we can obtain the convergence and convergence rates when the method (4.2) is terminated by either an *a priori* stopping rule or the discrepancy principle (1.5).

**Corollary 4.1** *For the linear ill-posed problem* (4.1) *in Hilbert spaces constrained by a closed convex set C, consider the dual projected Landweber iteration* (4.2) *with $\lambda_0 = 0$ and with noisy data $y^\delta$ satisfying $\|y^\delta - y\| \le \delta$.*

(i) *If $0 < \gamma \le 1/\|A\|^2$ then for the integer $n_\delta$ satisfying $n_\delta \to \infty$ and $\delta^2 n_\delta \to 0$ as $\delta \to 0$ there holds $\|x_{n_\delta} - x^\dagger\| \to 0$ as $\delta \to 0$. If in addition $x^\dagger$ satisfies the projected source condition*

$$x^\dagger = P_C((A^*A)^{\nu/2}\omega) \text{ for some } 0 < \nu \le 1 \text{ and } \omega \in X, \tag{4.3}$$

*then with the choice $n_\delta \sim \delta^{-\frac{2}{1+\nu}}$ we have $\|x_{n_\delta} - x^\dagger\| = O(\delta^{\frac{\nu}{1+\nu}})$.*

(ii) *If $\tau > 1$ and $\gamma > 0$ are chosen such that $1 - 1/\tau - \|A\|^2\gamma > 0$, then the discrepancy principle* (1.5) *defines a finite integer $n_\delta$ with $\|x_{n_\delta} - x^\dagger\| \to 0$ as $\delta \to 0$. If in addition $x^\dagger$ satisfies the projected source condition* (4.3)*, then $\|x_{n_\delta} - x^\dagger\| = O(\delta^{\frac{\nu}{1+\nu}})$.*

**Proof** According to Theorems 3.1, 3.4 and 3.7, it remains only to show that, under the projected source condition (4.3), $x^\dagger$ satisfies the variational source condition

$$\frac{1}{4}\|x - x^\dagger\|^2 \le \mathcal{R}(x) - \mathcal{R}(x^\dagger) + c_\nu\|\omega\|^{\frac{2}{1+\nu}}\|Ax - y\|^{\frac{2\nu}{1+\nu}} \tag{4.4}$$

for all $x \in \text{dom}(\mathcal{R}) = C$, where $c_\nu := 2^{-\frac{2\nu}{1+\nu}}(1 + \nu)(1 - \nu)^{\frac{1-\nu}{1+\nu}}$. To see this, note first that for any $x \in C$ there holds

$$\frac{1}{2}\|x - x^\dagger\|^2 - \mathcal{R}(x) + \mathcal{R}(x^\dagger) = \frac{1}{2}\|x - x^\dagger\|^2 - \frac{1}{2}\|x\|^2 + \frac{1}{2}\|x^\dagger\|^2$$
$$= \langle x^\dagger, x^\dagger - x \rangle.$$

By using $x^\dagger = P_C((A^*A)^{\nu/2}\omega)$ and the property of the projection $P_C$ we have

$$\langle (A^*A)^{\nu/2}\omega - x^\dagger, x - x^\dagger \rangle \le 0, \quad \forall x \in C$$

which implies that

$$\langle x^\dagger, x^\dagger - x \rangle \le \langle (A^*A)^{\nu/2}\omega, x^\dagger - x \rangle = \langle \omega, (A^*A)^{\nu/2}(x^\dagger - x) \rangle.$$

Therefore

$$\frac{1}{2}\|x - x^\dagger\|^2 - \mathcal{R}(x) + \mathcal{R}(x^\dagger) \le \langle \omega, (A^*A)^{\nu/2}(x^\dagger - x)\rangle$$
$$\le \|\omega\|\|(A^*A)^{\nu/2}(x^\dagger - x)\|.$$

By invoking the interpolation inequality [13] and the Young's inequality we can further obtain

$$\frac{1}{2}\|x - x^\dagger\|^2 - \mathcal{R}(x) + \mathcal{R}(x^\dagger) \le \|\omega\|\|x - x^\dagger\|^{1-\nu}\|A(x - x^\dagger)\|^\nu$$
$$\le c_\nu\|\omega\|^{\frac{2}{1+\nu}}\|Ax - y\|^{\frac{2\nu}{1+\nu}} + \frac{1}{4}\|x - x^\dagger\|^2$$

which shows (4.4). The proof is therefore complete. □

We remark that the projected source condition (4.3) with $\nu = 1$, i.e. $x^\dagger \in P_C(A^*Y)$, was first used in [38] to derive the convergence rate of Tikhonov regularization in Hilbert spaces with convex constraint.

The dual projected Landweber iteration (4.2) can be accelerated by Nesterov's acceleration strategy. As was derived in Sect. 3.3, the accelerated scheme takes the form

$$\hat{\lambda}_n = \lambda_n + \frac{n-1}{n+\alpha}(\lambda_n - \lambda_{n-1}), \qquad \hat{x}_n = P_C(A^*\hat{\lambda}_n),$$
$$\lambda_{n+1} = \hat{\lambda}_n - \gamma(A\hat{x}_n - y^\delta), \qquad x_{n+1} = P_C(A^*\lambda_{n+1}).$$

$$(4.5)$$

By noting that $\partial\mathcal{R}(x) = x + \partial\delta_C(x)$, it is easy to see that an element $\lambda^\dagger \in Y$ is such that $A^*\lambda^\dagger \in \partial\mathcal{R}(x^\dagger)$ if and only if $x^\dagger = P_C(A^*\lambda^\dagger)$. Therefore, by using Theorem 3.8, we can obtain the following convergence rate result of the method (4.5).

**Corollary 4.2** *For the problem* (4.1) *in Hilbert spaces constrained by a closed convex set $C$, consider the method* (4.5) *with $\lambda_0 = \lambda_{-1} = 0$. If $0 < \gamma \le 1/\|A\|^2$, $\alpha \ge 2$ and $x^\dagger \in P_C(A^*Y)$, then with the choice $n_\delta \sim \delta^{-1/2}$ we have*

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

*as $\delta \to 0$.*

### 4.2 An entropic dual gradient method

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and let $A : L^1(\Omega) \to Y$ be a bounded linear operator, where $Y$ is a Hilbert space. For an element $y \in Y$ in the range of $A$, we consider the equation $Ax = y$. We assume that the sought solution $x^\dagger$ is a probability density function, i.e. $x^\dagger \ge 0$ a.e. on $\Omega$ and $\int_\Omega x^\dagger = 1$. We may find such a solution

by considering the convex minimization problem

$$\min \left\{ \mathcal{R}(x) := f(x) + \delta_\Delta(x) : x \in L^1(\Omega) \text{ and } Ax = y \right\}, \tag{4.6}$$

where $\delta_\Delta$ denotes the indicator function of the closed convex set

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

in $L^1(\Omega)$ and $f$ denotes the negative of the Boltzmann-Shannon entropy, i.e.

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

where, here and below, $L^p_+(\Omega) := \{x \in L^p(\Omega) : x \geq 0 \text{ a.e. on } \Omega\}$ for each $1 \leq p \leq \infty$. The Boltzmann-Shannon entropy has been used in Tikhonov regularization as a stable functional to determine nonnegative solutions; see [1, 11, 14, 31] for instance.

In the following we summarize some useful properties of the negative of the Boltzmann-Shannon entropy $f$:

(i) $f$ is proper, lower semi-continuous and convex on $L^1(\Omega)$; see [1, 11].
(ii) $f$ is subdifferentiable at $x \in L^1(\Omega)$ if and only if $x \in L^\infty_+(\Omega)$ and is bounded away from zero, i.e.

$$\text{dom}(\partial f) = \{x \in L^\infty_+(\Omega) : x \geq \beta \text{ on } \Omega \text{ for some constant } \beta > 0\}.$$

Moreover for each $x \in \text{dom}(\partial f)$ there holds $\partial f(x) = \{1 + \log x\}$; see [4, Proposition 2.53].
(iii) By straightforward calculation one can see that for any $x \in \text{dom}(\partial f)$ and $\tilde{x} \in \text{dom}(f)$, the Bregman distance induced by $f$ is the Kullback-Leibler functional

$$D(\tilde{x}, x) := \int_\Omega \left( \tilde{x} \log \frac{\tilde{x}}{x} - \tilde{x} + x \right).$$

(iv) For any $x \in \text{dom}(\partial f)$ and $\tilde{x} \in \text{dom}(f)$ there holds (see [7, Lemma 2.2])

$$\|x - \tilde{x}\|^2_{L^1(\Omega)} \leq \left( \frac{4}{3} \|x\|_{L^1(\Omega)} + \frac{2}{3} \|\tilde{x}\|_{L^1(\Omega)} \right) D(\tilde{x}, x). \tag{4.7}$$

Based on these facts, we can see that the function $\mathcal{R}$ defined in (4.6) satisfies Assumption 1. In order to apply the dual gradient method (1.3) to solve (4.6), we need to determine the closed formula for the solution of the minimization problem involved in the algorithm. By the Karush-Kuhn-Tucker theory, it is easy to see that, for any $\ell \in L^\infty(\Omega)$, the unique minimizer of

$$\min \left\{ \int_\Omega (x \log x - \ell x) : x \geq 0 \text{ a.e. on } \Omega \text{ and } \int_\Omega x = 1 \right\}$$

is given by $\hat{x} = e^{\ell} / \int_{\Omega} e^{\ell}$. Therefore we can obtain from the algorithm (1.3) the following entropic dual gradient method

$$x_n = \frac{1}{\int_{\Omega} e^{A^* \lambda_n}} e^{A^* \lambda_n},$$

$$\lambda_{n+1} = \lambda_n - \gamma(Ax_n - y^{\delta}). \tag{4.8}$$

We have the following convergence result.

**Corollary 4.3** *For the convex problem* (4.6), *consider the entropic dual gradient method* (4.8) *with* $\lambda_0 = 0$ *and with noisy data* $y^{\delta}$ *satisfying* $\|y^{\delta} - y\| \leq \delta$.

(i) *If* $0 < \gamma \leq 1 / \|A\|^2$ *then for the integer* $n_{\delta}$ *satisfying* $n_{\delta} \to \infty$ *and* $\delta^2 n_{\delta} \to 0$ *as* $\delta \to 0$ *there holds* $\|x_{n_{\delta}} - x^{\dagger}\| \to 0$ *as* $\delta \to 0$. *If in addition* $x^{\dagger}$ *satisfies the source condition*

$$1 + \log x^{\dagger} = A^* \lambda^{\dagger} \quad \text{for some } \lambda^{\dagger} \in Y, \tag{4.9}$$

*then with the choice* $n_{\delta} \sim \delta^{-1}$ *we have* $\|x_{n_{\delta}} - x^{\dagger}\|_{L^1(\Omega)} = O(\delta^{1/2})$.

(ii) *If* $\tau > 1$ *and* $\gamma > 0$ *are chosen such that* $1 - 1/\tau - \|A\|^2 \gamma > 0$, *then the discrepancy principle* (1.5) *defines a finite integer* $n_{\delta}$ *with* $\|x_{n_{\delta}} - x^{\dagger}\|_{L^1(\Omega)} \to 0$ *as* $\delta \to 0$. *If in addition* $x^{\dagger}$ *satisfies the source condition* (4.9), *then* $\|x_{n_{\delta}} - x^{\dagger}\|_{L^1(\Omega)} = O(\delta^{1/2})$.

**Proof** Under (4.9) there holds $A^* \lambda^{\dagger} \in \partial f(x^{\dagger})$. Therefore, by using (4.7), we have for any $x \in \text{dom}(\mathcal{R})$ that

$$\frac{1}{2} \|x - x^{\dagger}\|^2_{L^1(\Omega)} \leq D(x, x^{\dagger}) = f(x) - f(x^{\dagger}) - \langle A^* \lambda^{\dagger}, x - x^{\dagger} \rangle$$

$$= \mathcal{R}(x) - \mathcal{R}(x^{\dagger}) - \langle \lambda^{\dagger}, Ax - y \rangle$$

$$\leq \mathcal{R}(x) - \mathcal{R}(x^{\dagger}) + \|\lambda^{\dagger}\| \|Ax - y\|,$$

where we used $\mathcal{R}(x) = f(x)$ and $\int_{\Omega} x = 1$ for $x \in \text{dom}(\mathcal{R})$. Thus $x^{\dagger}$ satisfies the varaitional source condition specified in Assumption 2 with $\mathcal{E}^{\dagger}(x) = \frac{1}{2} \|x - x^{\dagger}\|^2_{L^1(\Omega)}$, $M = \|\lambda^{\dagger}\|$ and $q = 1$. Now we can complete the proof by applying Theorems 3.1, 3.7 and Corollary 3.5 to the method (4.8). □

The source condition (4.9) has been used in [11, 14] in which one may find further discussions. We would like to mention that an entropic Landweber method of the form

$$x_{n+1} = \frac{x_n e^{\gamma A^* (y^{\delta} - Ax_n)}}{\int_{\Omega} x_n e^{\gamma A^* (y^{\delta} - Ax_n)}} \tag{4.10}$$

has been proposed and studied in the recent paper [10] in which weak convergence in $L^1(\Omega)$ is proved without relying on source conditions and, under the source condition (4.9), an error estimate is derived when the method is terminated by an *a priori* stopping

rule. Our method (4.8) is different from (4.10) due to its primal-dual nature. As stated in Corollary 4.3, our method (4.8) enjoys nicer convergence properties: it admits strong convergence in $L^1(\Omega)$ in general and, when the source condition (4.9) is satisfied, an error estimate can be derived when the method is terminated by either an *a priori* stopping rule or the discrepancy principle.

Applying Nesterov's acceleration strategy, we can accelerate the entropic dual gradient method (4.8) by the following scheme

$$\hat{\lambda}_n = \lambda_n + \frac{n-1}{n+\alpha}(\lambda_n - \lambda_{n-1}), \qquad \hat{x}_n = \frac{1}{\int_\Omega e^{A^*\hat{\lambda}_n}} e^{A^*\hat{\lambda}_n},$$
$$\lambda_{n+1} = \hat{\lambda}_n - \gamma(A\hat{x}_n - y^\delta), \qquad x_{n+1} = \frac{1}{\int_\Omega e^{A^*\lambda_{n+1}}} e^{A^*\lambda_{n+1}}. \tag{4.11}$$

By using Theorem 3.8 we can obtain the following convergence rate result on the method (4.11) with noisy data.

**Corollary 4.4** *For the minimization problem* (4.6), *consider the method* (4.11) *with* $\lambda_0 = \lambda_{-1} = 0$. *If* $0 < \gamma \le 1/\|A\|^2$, $\alpha \ge 2$ *and* $x^\dagger$ *satisfies the source condition* (4.9), *then with the choice* $n_\delta \sim \delta^{-1/2}$ *we have*

$$\|x_{n_\delta} - x^\dagger\|_{L^1(\Omega)} = O(\delta^{1/2})$$

*as* $\delta \to 0$.

# 5 Conclusion

Due to its simplicity and relatively small complexity per iteration, Landweber iteration has received extensive attention in the inverse problems community. In recent years, Landweber iteration has been extended to solve inverse problems in Banach spaces with general uniformly convex regularization terms and various convergence properties have been established. However, except for the linear and nonlinear Landweber iteration in Hilbert spaces, the convergence rate in general is missing from the existing convergence theory.

This paper attempts to fill in this gap by providing a novel technique to derive convergence rates for a class of Landweber type methods. We considered a class of ill-posed problems defined by a bounded linear operator from a Banach space to a Hilbert space and used a strongly convex regularization functional to select the sought solution. The dual problem turns out to have a smooth objective function and thus can be solved by the usual gradient method. The resulting method is called a dual gradient method which is a special case of the Landweber type method in Banach spaces. Applying gradient methods to the dual problem allows us to interpret the method in a new perspective which enables us to use tools from convex analysis and optimization to carry out the analysis. We have actually obtained the convergence and convergence rates of the dual gradient method when it is terminated by either an *a priori* stopping rule or the discrepancy principle. Furthermore, by applying Nesetrov's acceleration

strategy to the dual problem we proposed an accelerated dual gradient method and established a convergence rate result under an *a priori* stopping rule. We also discussed some applications, in particular, as a direct application of our convergence theory, we provided a rather complete analysis of the dual projected Landweber iteration of Eicke for which only preliminary result is available in the existing literature.

There are a few of questions which might be interesting for future development.

(i) We established convergence rate results for the dual gradient method (1.3) which require $A$ to be a bounded linear operator and $Y$ a Hilbert space. Is it possible to establish a general convergence rate result for Landweber iteration for solving linear as well as nonlinear ill-posed problems in Banach spaces?

(ii) For the dual gradient method (1.3), its analysis under the *a priori* stopping rule allows to take the step-size as $0 < \gamma \le 1/L$, while the analysis under the discrepancy principle (1.5) requires $\tau > 1$ and $\gamma > 0$ to satisfy $1 - 1/\tau^2 - L\gamma > 0$ which means either $\tau$ has to be large or $\gamma$ has to be small. Is it possible to develop a convergence theory of the discrepancy principle under merely the conditions $\tau > 1$ and $0 < \gamma \le 1/L$?

(iii) In Sect. 3.3 we considered the accelerated dual gradient method (3.18) and established a convergence rate result under an *a priori* stopping rule. Is it possible to establish the convergence and convergence rate result of (3.18) under the discrepancy principle?

# References

1. Amato, U., Hughes, W.: Maximum entropy regularization of Fredholm integral equations of the first kind. Inverse Probl. **7**, 793–808 (1991)
2. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov accelerated forward-backward method is actually faster than $O(1/k^2)$. SIAM J. Optim. **26**, 1824–1834 (2016)
3. Aujol, J.F., Dossal, Ch.: Stability of over-relaxations for the forward-backward algorithm, application to FISTA. SIAM J. Optim. **25**(4), 2408–2433 (2015)
4. Barbu, V., Precupanu, T.: Convexity and Optimization in Banach Spaces. Springer Monographs in Mathematics, 4th edn. Springer, Berlin (2012)
5. Beck, A.: First-Order Methods in Optimization, MOS-SIAM Series on Optimization 25, SIAM, (2017)
6. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**, 183–202 (2009)
7. Borwein, J.M., Lewis, C.S.: Convergence of best entropy estimates. SIAM J. Optim. **1**, 191–205 (1991)

8. Borwein, J.M., Zhu, Q.J.: Techniques of variational analysis, CMS Books in Mathematics/Ouvrages de Mathematiques de la SMC, 20. Springer, New York (2005)
9. Boţ, R., Hein, T.: Iterative regularization with a geeral penalty term - theory and applications to $L^1$ and TV regularization. Inverse Probl. **28**, 104010 (2012)
10. Burger, M., Resmerita, E., Benning, M.: An entropic Landweber method for linear ill-posed problems. Inverse Probl. **36**(1), 015009 (2020)
11. Eggermont, P.P.B.: Maximum entropy regularization for Fredholm integral equations of the first kind. SIAM J. Math. Anal. **24**(6), 1557–1576 (1993)
12. Eicke, B.: Iteration methods for convexly constrained ill-posed problems in Hilbert space. Numer. Funct. Anal. Optim. **13**, 413–429 (1992)
13. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
14. Engl, H.W., Landl, G.: Convergence rates for maximum entropy regularization. SIAM J. Numer. Anal. **30**, 1509–1536 (1993)
15. Flemming, J.: Existence of variational source conditions for nonlinear inverse problems in Banach spaces. J. Inverse Ill-posed Probl. **26**(2), 277–286 (2018)
16. Flemming, J., Gerth, D.: Injectivity and weak*-to-weak continuity suffices for convergence rates in $\ell^1$ regularization. J. Inverse Ill-posed Probl. **26**(1), 85–94 (2018)
17. Frick, K., Grasmair, M.: Regularization of linear ill-posed problems by the augmented Lagrangian method and variational inequalities. Inverse Probl. **28**(10), 104005 (2012)
18. Frick, K., Lorenz, D., Resmerita, E.: Morozov's principle for the augmented Lagrangian method applied to linear inverse problems. Multiscale Model. Simul. **9**(4), 1528–1548 (2011)
19. Frick, K., Scherzer, O.: Regularization of ill-posed linear equations by the non-stationary augmented Lagrangian method. J. Integral Equ. Appl. **22**(2), 217–257 (2010)
20. Grasmair, M.: Non-convex sparse regularisation. J. Math. Anal. Appl. **365**(1), 19–28 (2010)
21. Hanke, M.: Accelerated Landweber iterations for the solution of ill-posed equations. Numer. Math. **60**, 341–373 (1991)
22. Hanke, M., Neubauer, A., Scherzer, O.: A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. Numer. Math. **72**, 21–37 (1995)
23. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Probl. **23**, 987–1010 (2007)
24. Hofmann, B., Mathé, P.: Parameter choice under variational inequalities. Inverse Probl. **28**, 104006 (2012)
25. Hohage, T., Weidling, F.: Verification of a variational source condition for acoustic inverse medium scattering problems. Inverse Probl. **31**(7), 075006 (2015)
26. Hohage, T., Weidling, F.: Characterizations of variational source conditions, converse results, and maxisets of spectral regularization methods. SIAM J. Numer. Anal. **55**(2), 598–620 (2017)
27. Hubmer, S., Ramlau, R.: Convergence analysis of a two-point gradient method for nonlinear ill-posed problems. Inverse Probl. **33**(9), 095004 (2017)
28. Hubmer, S., Ramlau, R.: Nesterov's accelerated gradient method for nonlinear ill-posed problems with a locally convex residual functional. Inverse Probl. **34**, 095003 (2018)
29. Jin, Q.: Landweber-Kaczmarz method in Banach spaces with inexact inner solvers. Inverse Probl. **32**, 104005 (2016)
30. Jin, Q.: On a heuristic stopping rule for the regularization of inverse problems by the augmented Lagrangian method. Numer. Math. **136**(4), 973992 (2017)
31. Jin, Q., Hou, Z.: Maximum entropy method for solving nonlinear ill-posed problems. Chin. Sci. Bull. **41**, 1589–1593 (1996)
32. Jin, Q., Stals, L.: Nonstationary iterated Tikhonov regularization for ill-posed problems in Banach spaces. Inverse Probl. **28**(10), 104011 (2012)
33. Jin, Q., Wang, W.: Landweber iteration of Kaczmarz type with general non-smooth convex penalty functionals. Inverse Probl. **29**, 085011 (2013)
34. Kaltenbacher, B., Schöpfer, F., Schuster, T.: Convergence of some iterative methods for the regularization of nonlinear ill-posed problems in Banach spaces. Inverse Probl. **25**(6), 065003 (2009)
35. Kindermann, S.: Optimal-order convergence of Nesterov acceleration for linear ill-posed problems. Inverse Prob. **37**(6), 065002 (2021)
36. Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001)
37. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Doklady **27**, 372–376 (1983)

38. Neubauer, A.: Tikhonov-regularization of ill-posed linear operator equations on closed convex sets. J. Approx. Theory **53**(3), 304–320 (1988)
39. Neubauer, A.: On Nesterov acceleration for Landweber iteration of linear ill-posed problems. J. Inverse Ill-Posed Probl. **25**(3), 381–390 (2017)
40. Real, R., Jin, Q.: A revisit on Landweber iteration. Inverse Probl. **36**(7), 075011 (2020)
41. Schöpfer, F., Louis, A.K., Schuster, T.: Nonlinear iterative methods for linear ill-posed problems in Banach spaces. Inverse probl. **22**, 311–329 (2006)
42. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.S.: Regularization Methods in Banach Spaces. Radon Series on Computational and Applied Mathematics, Walter de Gruyter, Berlin (2012)
43. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. **29**(1), 119–138 (1991)
44. Zălinscu, C.: Convex Analysis in General Vector Spaces. World Scientific Publishing Co., Inc., River Edge (2002)
45. Zhong, M., Wang, W., Jin, Q.: Regularization of inverse problems by two-point gradient methods in Banach spaces. Numer. Math. **143**(3), 713–747 (2019)