# Acceleration of sequentialsubspace optimization in Banach spaces by orthogonal search directions. [*]

Frederik Heber[1] Frank Schöpfer[2], Thomas Schuster[3]

IfM Preprint No. 2018-03
September 2018

Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
D-26111 Oldenburg, Germany

[*] Published in *Journal of Computational and Applied Mathematics, 345:1-?22, 2019.*

[1] heber@math.uni-sb.de

[2] frank.schoepfer@uni-oldenburg.de

[3] thomas.schuster@num.uni-sb.de

# Acceleration of sequential subspace optimization in Banach spaces by orthogonal search directions

Frederik Heber[a,*], Frank Schöpfer[b], Thomas Schuster[a,*]

[a] *Department of Mathematics, Saarland University, 66041 Saarbrücken, Germany*
[b] *Department of Mathematics, Carl von Ossietzky University Oldenburg, 26129 Oldenburg, Germany*

## Abstract

A standard solution technique for linear operator equations of first kind is the Landweber scheme which is an iterative method that uses the negative gradient of the current residual as search direction, which is then also called the Landweber direction. Though this method proves to be stable with respect to noisy data, it is known to be numerically slow for problems in Hilbert spaces and this behavior shows to be even worse in some Banach space settings. This is why the idea came up to use several search directions instead of the Landweber direction only which has led to the development of Sequential Subspace Optimization (SESOP) methods. This idea is related to the famous Conjugate Gradient (CG) techniques that are known to be amongst the most effective methods to solve linear equations in Hilbert spaces. Since CG methods decisively make use of the inner product structure, they have been inherently restricted to Hilbert spaces so far. SESOP methods in Banach spaces do not share the conjugacy property with CG methods. In this article we use the concept of generalized orthogonality in Banach spaces and apply metric projections to orthogonalize the current Landweber direction with respect to the search space of the last iteration. This leads to an accelerated SESOP method which is confirmed by various numerical

[*] Corresponding author

*Email addresses:* `heber@math.uni-sb.de` (Frederik Heber), `frank.schoepfer@uni-oldenburg.de` (Frank Schöpfer), `thomas.schuster@num.uni-sb.de` (Thomas Schuster)

*URL:* `www.num.uni-sb.de/schuster` (Thomas Schuster)

experiments. Moreover, in Hilbert spaces our method coincides with the Conjugate Gradient Normal Error (CGNE) or Craig's method applied to the normal equation. We prove weak convergence to the exact solution. Furthermore we perform a couple of numerical tests on a linear problem involving a random matrix and on the problem of 2D computerized tomography where we use different $\ell_p$-spaces. In all experiments the orthogonalization of the search space shows superior convergence properties compared to standard SESOP. This especially holds for $p$ close to 1. Letting $p \to 2$ the more we recover the conjugacy property for the search directions and the more the convergence behaves independently of the size of the search space.

*Key words:* metric projection, Bregman distance, Banach space, sequential subspace optimization, CGNE, Bregman projection, computerized tomography
*2010 MSC:* 46B20, 65F10, 65K10, 65R32

## 1. Introduction

We consider two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$ and a bounded, linear operator

$$\mathbf{A} : \mathcal{X} \to \mathcal{Y}. \tag{1}$$

Our aim is to iteratively solve the inverse problem

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{2}$$

for given $\mathbf{y} \in \mathcal{Y}$. The space $\mathcal{X}$ is assumed to be smooth and uniformly convex and hence $\mathcal{X}$ is reflexive and has a strictly convex and uniformly smooth dual $\mathcal{X}^*$. The space $\mathcal{Y}$ can be arbitrary. Problem (2) may be ill-posed and thus not suitable for direct inversion of the operator. Hence, a regularization scheme is required to obtain a stable solution.

In [1] the Landweber method, which has been thoroughly investigated in Hilbert spaces; see e.g. [2], has been extended to this Banach space setting. It is a steepest descent method, see [3], since it uses as search direction the gradient of the quadratic residual $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 / 2$ at the current iterate $\mathbf{x}_n$. This direction, $\mathbf{A}^*(\mathbf{A}\mathbf{x}_n - \mathbf{y})$, is also called the *Landweber direction.* The Landweber iteration converges under certain assumptions to the minimum-norm solution [2, 1]. However, the method usually suffers from tremendously slow convergence.

To speed up convergence the authors in [4] developed the idea of using finite dimensional subspaces spanned by search directions in each iteration, where the Landweber direction is included to guarantee convergence. This idea obviously was inspired by Conjugate Gradient (CG) methods in Hilbert spaces; see e.g. [5, 6]. In every iteration a Bregman projection onto these subspaces has to be computed. This is the Sequential Subspace Optimization (SESOP) method which goes back to Narkiss and Zibulevsky [7] and turns into a regularization method if the discrepancy principle as parameter choice rule is used. In general only weak convergence to the minimum-norm solution can be proven, see [7]. Taking specific search directions into account also strong convergence has been proven, see [8, Prop. 1]. The method has furthermore been extended to nonlinear inverse problems in Hilbert spaces [9]. However, the set of search directions is still not optimal. In the thesis of [10] the search directions are further modified in the notion of maximizing their pairwise orthogonality. In this article we give its detailed derivation, which also connects the SESOP techniques with the family of CG methods in Hilbert spaces. More specifically we prove that the SESOP with orthogonalized search directions in Banach spaces boils down to the Conjugate Gradient Normal Error (CGNE) method (also known as *Craig's method*) [11], if $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. The orthogonalization is done by using appropriate metric projections. We furthermore prove weak convergence to the Bregman projection of the initial value onto the solution manifold $\{\mathbf{x} \in \mathcal{X} : \mathbf{A}\mathbf{x} = \mathbf{y}\}$.

*Outline.* In Section 2 we recollect general properties of duality mappings, uniformly smooth Banach spaces, metric and Bregman projections. Starting with the SESOP method (Section 3.1) we develop in Section 3 orthogonalized search spaces to get an accelerated solver where we use and define the concept of *generalized orthogonality* in Banach spaces due to Alber [12] (Section 3.2 and 3.3).

The orthogonalized search directions are connected to the Landweber directions of the last $N$ iterates. We prove weak convergence of the method (Section 3.4). Section 4 finally consists of a thorough numerical validation of our method where we first use a linear system with a random matrix and then check the performance on the problem of 2D computerized tomography using $\ell_p$-settings for different values of $p$. We conclude that the orthogonalization of the search directions leads to a significant acceleration in computing time compared to the experiments with SESOP in [4].

## 2. Preliminaries

Throughout the paper let $\mathcal{X}$ and $\mathcal{Y}$ be real Banach spaces with duals $\mathcal{X}^*$ and $\mathcal{Y}^*$, respectively. The space $\mathcal{X}$ is assumed to be smooth and uniformly convex with a sequentially weak-to-weak continuous duality mapping (see Subsection 2.1). The space $\mathcal{Y}$ is arbitrary. The norms will be denoted by $\|.\|_{\mathcal{X}}$ and $\|.\|_{\mathcal{X}^*}$, respectively. For $\mathbf{x} \in \mathcal{X}$ and $\mathbf{x}^* \in \mathcal{X}^*$, we write $\langle \mathbf{x}, \mathbf{x}^* \rangle = \langle \mathbf{x}^*, \mathbf{x} \rangle = \mathbf{x}^*(\mathbf{x})$ for the dual pairing in $\mathcal{X}^* \times \mathcal{X}$ and define $\langle \mathbf{y}^*, \mathbf{y} \rangle_{\mathcal{Y}^* \times \mathcal{Y}}$ accordingly. We omit subindices whenever it is clear which norm or dual pairing is meant. By $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ we designate the space consisting of all bounded, linear operators $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ and write $\mathbf{A}^*$ for its adjoint operator $\mathbf{A}^* \in \mathcal{L}(\mathcal{Y}^*, \mathcal{X}^*)$. We denote by $\operatorname{ran}(\mathbf{A})$ the range and by $\operatorname{nul}(\mathbf{A})$ the null space of $\mathbf{A}$.

For real numbers $a$, $b$, we write

$$a \vee b = \max\{a, b\}, \quad a \wedge b = \min\{a, b\}.$$

Also, let $p, p^*, r, r^* \in (1, \infty)$ be conjugate exponents such that

$$\frac{1}{p} + \frac{1}{p^*} = 1 \quad \text{and} \quad \frac{1}{r} + \frac{1}{r^*} = 1.$$

In the following subsections we recollect some important concepts of Banach space theory that are useful to establish the method and for convergence analysis.

### *2.1. Duality Mappings*

We recall the definition of the subdifferential of a convex function and duality mappings in Banach spaces and some of their properties, all of which can be found in the comprehensive book [13].

**Definition 1** (Subdifferential)**.** Let $f : \mathcal{X} \to \mathbb{R}$ be convex. The *subdifferential* $\partial f(\mathbf{x})$ of $f$ at $\mathbf{x} \in \mathcal{X}$ is given as

$$\partial f(\mathbf{x}) = \Big\{ \mathbf{x}^* \in \mathcal{X}^* : f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{x}^*, \mathbf{y} - \mathbf{x} \rangle \quad \text{for all } \mathbf{y} \in \mathcal{X} \Big\}.$$

**Definition 2** (Duality Mapping)**.** Let $\mathcal{X}$ be a Banach space. The mapping $J_p : \mathcal{X} \to 2^{\mathcal{X}^*}$ defined by

$$J_p(\mathbf{x}) = \{\mathbf{x}^* \in \mathcal{X}^* \mid \langle \mathbf{x}^*, \mathbf{x} \rangle = \|\mathbf{x}\|^p \,, \|\mathbf{x}^*\| = \|\mathbf{x}\|^{p-1}\} \tag{3}$$

is the *duality mapping* of $\mathcal{X}$ with gauge function $t \mapsto t^{p-1}$. Accordingly $J^*_{p^*}$ denotes the duality mapping on $\mathcal{X}^*$ with power $p^*$.

By the Theorem of Asplund, see [13, Thm. 4.4], we have

$$J_p(\mathbf{x}) = \partial \left( \tfrac{1}{p} \|\mathbf{x}\|^p \right). \tag{4}$$

**Definition 3** (Smooth Banach spaces)**.** We call the Banach space $\mathcal{X}$ *smooth*, if for every $\mathbf{x} \in \mathcal{X}$ with $\mathbf{x} \neq 0$ there is a unique element $\mathbf{x}^* \in \mathcal{X}^*$ such that $\|\mathbf{x}^*\| = 1$ and $\langle \mathbf{x}^*, \mathbf{x} \rangle = \|\mathbf{x}\|$.

The reason for calling a Banach space to be smooth is that its norm is Gâteaux differentiable. In this case the duality mapping $J_p$ is single-valued.

**Proposition 1** ([13, Thm. 3.5, 4.5])**.** For a Banach space $\mathcal{X}$ we have the equivalences:

a) The Banach space $\mathcal{X}$ is smooth.

b) The norm $\|\cdot\|_{\mathcal{X}}$ is Gâteaux differentiable on $\mathcal{X} \setminus \{0\}$.

c) The duality mapping $J_p$ is single-valued.

**Theorem 1** ([14, Th. 2.53])**.** If the Banach space $\mathcal{X}$ is smooth, strictly convex and reflexive, then $J_p$ is single-valued, norm-to-weak continuous, bijective, and the duality mapping $J^*_{p^*}$ is single-valued, too, and satisfies

$$J^*_{p^*}\big(J_p(\mathbf{x})\big) = \mathbf{x} \qquad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{5}$$

For this situation Figure 1 illustrates the connection between different Banach spaces, their duals and the respective duality mappings. The relation (5) is essential for the construction of Landweber and SESOP methods in Banach spaces.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\mathbf{A}"] \arrow[d, bend right, "J_{\mathcal{X}}"'] & \mathcal{Y} \arrow[d, "J_{\mathcal{Y}}"] \\
\mathcal{X}^* \arrow[u, bend right, "J_{\mathcal{X}^*}"'] & \mathcal{Y}^* \arrow[l, "\mathbf{A}^*"']
\end{tikzcd}$$

**Figure 1:** Relations between the Banach spaces and dual spaces involved in the inverse problem $\mathbf{Ax} = \mathbf{y}$.

### *2.2. Uniform smoothness*

**Definition 4** (Uniformly smooth, [15, Def. II.1.e.1])**.** Let $\mathcal{X}$ be a Banach space with $\dim \mathcal{X} \geq 2$.

(a) The *modulus of smoothness* is defined by

$$\rho_{\mathcal{X}}(\tau) = \tfrac{1}{2} \sup_{\|\mathbf{x}\|=1, \|\mathbf{y}\|=1} \big( \|\mathbf{x} + \tau\mathbf{y}\| + \|\mathbf{x} - \tau\mathbf{y}\| - 2\big), \quad \tau > 0. \tag{6}$$

(b) $\mathcal{X}$ is said to be uniformly smooth if

$$\lim_{\tau \to \infty} \frac{\rho_{\mathcal{X}}(\tau)}{\tau} = 0. \tag{7}$$

Note that $L_p$-spaces and $\ell_p$-spaces with $p \in (1, \infty)$ are uniformly convex, see [16], and also uniformly smooth, see [17, Lemma 6.7], see also [18, p. 63]. On the other hand, $\ell_1$ and $\ell_\infty$ are not reflexive[1], see [19, Sect. VIII. 5], and hence are neither uniformly smooth, nor uniformly convex.

In order to prove convergence we will heavily rely on the geometrical characteristics of Banach spaces. Essentially, we need to estimate $\|\mathbf{x} - \mathbf{y}\|$ by the norms $\|\mathbf{x}\|, \|\mathbf{y}\|$. To this end the well-known Xu-Roach inequalities are important, one of which we recall here for convenience.

**Theorem 2** ([20, Theorem 2, Remark 4])**.** If $\mathcal{X}$ is uniformly smooth, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$\|\mathbf{x} - \mathbf{y}\|^p \leq \|\mathbf{x}\|^p - p\langle J_p(\mathbf{x}), \mathbf{y}\rangle + \widetilde{\sigma}_p(\mathbf{x}, \mathbf{y}) \tag{8}$$

with

$$\widetilde{\sigma}_p(\mathbf{x}, \mathbf{y}) = p\, G_{p^*} \int_0^1 \frac{(\|\mathbf{x} - t\mathbf{y}\| \vee \|\mathbf{x}\|)^p}{t} \rho_{\mathcal{X}} \left( \frac{t\,\|\mathbf{y}\|}{\|\mathbf{x} - t\mathbf{y}\| \vee \|\mathbf{x}\|} \right) dt. \tag{9}$$

and a constant $G_{p^*} > 0$.

We note that functions $\widetilde{\sigma}_{p^*}$ defined on $\mathcal{X}^* \times \mathcal{X}^*$ and a constant $G_p > 0$ are defined in the same way as $\widetilde{\sigma}_p$ and $G_{p^*}$ with $p$ and $p^*$ switched. We further state a lemma on an upper bound of the function $\widetilde{\sigma}_{p^*}$, that we need later in the convergence proof.

**Lemma 1** (Upper bound on $\widetilde{\sigma}_{p^*}$,[10, Proof of Prop. 2.39])**.** Let $\mathcal{X}^*$ be a uniformly smooth Banach space with duality mapping $J^*_{p^*}$. If $0 \neq \mathbf{x} \in \mathcal{X}$, $0 \neq \mathbf{A} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $0 \neq \mathbf{y}^* \in \mathcal{Y}^*$ with an arbitrary Banach space $\mathcal{Y}$ are given and $\mu > 0$ is defined by

$$\mu := \frac{\tau}{\|\mathbf{A}\|} \frac{\|\mathbf{x}\|^{p-1}}{\|\mathbf{y}^*\|} \quad \text{for some } \tau \in (0, 1], \tag{10}$$

then the following estimate is valid:

$$\tfrac{1}{p^*} \widetilde{\sigma}_{p^*}(J_{p^*}(\mathbf{x}), \mu \mathbf{A}^* \mathbf{y}^*) \leq 2^{p^*} G_p \|\mathbf{x}\|^p \rho_{\mathcal{X}^*}(\tau), \tag{11}$$

[1] Note that finite-dimensional spaces are always reflexive.

where $\rho_{\mathcal{X}^*}$ is the modulus of smoothness of $\mathcal{X}^*$.

*Proof.* From an analogue definition (9) for $\widetilde{\sigma}_{p^*}(J_{p^*}(\mathbf{x}), \mathbf{A}^*\mathbf{y}^*)$ we can estimate using (10)

$$\|J_{p^*}(\mathbf{x}) - t\mu\mathbf{A}^*\mathbf{y}^*\| \leq \|\mathbf{x}\|^{p-1} + \mu\,\|\mathbf{A}\|\,\|\mathbf{y}^*\| \leq 2\,\|\mathbf{x}\|^{p-1}$$

and

$$\|\mathbf{x}\|^{p-1} \leq \|J_{p^*}(\mathbf{x}) - t\mu\mathbf{A}^*\mathbf{y}^*\| \vee \|J_{p^*}(\mathbf{x})\| \leq 2\,\|\mathbf{x}\|^{p-1}\,.$$

As the modulus of smoothness $\rho_{\mathcal{X}^*}$ is non-decreasing, see [18, Prop. 1.e.5], and by (4) and (10), we obtain

$$\rho_{\mathcal{X}^*}\left(\frac{t\mu\,\|\mathbf{A}^*\mathbf{y}^*\|}{\|J_{p^*}(\mathbf{x}) - \mu\mathbf{A}^*\mathbf{y}^*\| \vee \|J_{p^*}(\mathbf{x})\|}\right) \leq \rho_{\mathcal{X}^*}\left(\frac{t\mu\,\|\mathbf{A}^*\mathbf{y}^*\|}{\|\mathbf{x}\|^{p-1}}\right) \leq \rho_{\mathcal{X}^*}(t\tau).$$

We finally arrive at the desired estimate,

$$\begin{aligned}\tfrac{1}{p^*}\widetilde{\sigma}_{p^*}(J_{p^*}(\mathbf{x}), \mu\mathbf{A}^*\mathbf{y}^*) &\leq 2^{p^*} G_p\,\|\mathbf{x}\|^p \int_0^1 \frac{\rho_{\mathcal{X}^*}(t\tau)}{t}dt \\ &= 2^{p^*} G_p\,\|\mathbf{x}\|^p \int_0^\tau \frac{\rho_{\mathcal{X}^*}(t)}{t}dt \\ &\leq 2^{p^*} G_p\,\|\mathbf{x}\|^p\,\rho_{\mathcal{X}^*}(\tau)\end{aligned}$$

as the function $\tau \to \frac{\rho_{\mathcal{X}^*}(\tau)}{\tau}$ is non-decreasing, see [21, Cor. 2.8] □

### *2.3. Metric and Bregman projections*

In this subsection we summarize essential properties of metric and Bregman projections. In this section we always assume that $\mathcal{C} \subset \mathcal{X}$ is a non-empty, closed and convex subset.

**Definition 5** (Metric Projection)**.** The *metric projection* $P$ of $\mathbf{x} \in \mathcal{X}$ onto $\mathcal{C}$ is the unique element $P_{\mathcal{C}}(\mathbf{x}) \in \mathcal{C}$ such that

$$\|\mathbf{x} - P_{\mathcal{C}}(\mathbf{x})\| = \min_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x} - \mathbf{y}\|\,. \tag{12}$$

Let us recall the Bregman distance in the context of generalized distance functions, see also [22, Sect. 2.1].

**Definition 6** (Bregman distance)**.** For a convex function $f : \mathcal{X} \to \mathbb{R}$ and $\xi \in \partial f(x) \subset \mathcal{X}^*$ the *Bregman distance* $\Delta_f : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ associated with $\xi$ is defined as

$$\Delta_f(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) - f(\mathbf{x}) - \langle \xi, \mathbf{y} - \mathbf{x} \rangle, \quad \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{13}$$

In our article we especially consider Bregman distances of functions $f_p(\mathbf{x}) = \frac{1}{p} \left\| \mathbf{x} \right\|^p$ with $\partial f_p(x) = J_p$. The corresponding Bregman distance is then denoted by $\Delta_p := \Delta_{f_p}$. A useful identity [4] for $\Delta_p$ is given by

$$\Delta_p(\mathbf{x}, \mathbf{y}) = \frac{1}{p^*} \left\| \mathbf{x} \right\|^p - \langle J_p(\mathbf{x}), \mathbf{y} \rangle + \frac{1}{p} \left\| \mathbf{y} \right\|^p. \tag{14}$$

We collect some important properties of the Bregman distance.

**Proposition 2** (Properties of Bregman Distances,[1, Theorem 2.12])**.** For all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and sequences $\{\mathbf{x_n}\}_n$ in $\mathcal{X}$ we have:

(a) $\Delta_p(\mathbf{x}, \mathbf{y}) \geq 0$ and $\Delta_p(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$.

(b) $\lim_{\|x_n\| \to \infty} \Delta_p(\mathbf{x_n}, \mathbf{x}) = \infty$, i. e. the sequence $\{\mathbf{x_n}\}_n$ remains bounded if the sequence $\left\{\Delta_p(\mathbf{x_n}, \mathbf{x})\right\}_n$ is bounded.

(c) $\Delta_p$ is continuous in both arguments. It is strictly convex and Gâteaux differentiable with respect to the second variable with $\partial_{\mathbf{y}} \Delta_p(\mathbf{x}, \mathbf{y}) = J_p(\mathbf{y}) - J_p(\mathbf{x})$.

It is easy to see that in Hilbert spaces metric distance and Bregman distance coincide. For $p = 2$ we have $\Delta_2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$.

We finally introduce the concept of the Bregman projection. Bregman projections minimize the Bregman distance with respect to a given closed, convex, non-empty set.

**Definition 7** (Bregman Projection)**.** The *Bregman projection* of $\mathbf{x} \in \mathcal{X}$ onto $\mathcal{C}$ with respect to the function $f_p(\mathbf{x}) = \frac{1}{p} \left\| \mathbf{x} \right\|_{\mathcal{X}}^p$ is the unique element $\Pi_{\mathcal{C}}^p(\mathbf{x}) \in \mathcal{C}$ such that

$$\Delta_p\left(\mathbf{x}, \Pi_{\mathcal{C}}^p(\mathbf{x})\right) = \min_{\mathbf{y} \in \mathcal{X}} \Delta_p(\mathbf{x}, \mathbf{y}). \tag{15}$$

We state some important relationships between Bregman and metric projections.

**Proposition 3** ([4, Prop. 3.6])**.**

(a) The Bregman projection and the metric projection are related via

$$P_{\mathcal{C}}(\mathbf{x}) - \mathbf{x} = \Pi^p_{\mathcal{C}-\mathbf{x}}(0) \quad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{16}$$

Especially we have $P_{\mathcal{C}}(0) = \Pi^p_{\mathcal{C}}(0)$.

(b) The metric projection satisfies the translation property

$$P_{\mathbf{y}+\mathcal{C}}(\mathbf{x}) = \mathbf{y} + P_{\mathcal{C}}(\mathbf{x}-\mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{17}$$

Property (b) indeed distinguishes the metric from the Bregman projection since if we had $\Pi^p_{\mathbf{y}+\mathcal{C}}(x) = \mathbf{y} + \Pi^p_{\mathcal{C}}(\mathbf{x}-\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, then this would already imply their equivalence, $\Pi^p_{\mathbf{y}+\mathcal{C}}(\mathbf{x}) = P_{\mathbf{y}+\mathcal{C}}(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

## 3. Methods

We now discuss the sequential subspace methods for solving problem (2). As a starting point we recall the sequential subspace optimization (SESOP) method as it is outlined in [4, 8] (Section 3.1). Subsequently we develop the concept of generalized orthogonality in Banach spaces (Section 3.2) and use this to orthogonalize the space of search directions that is used in the SESOP method (Section 3.3). This leads to an accelerated version, whose convergence analysis is outlined in Section 3.4.

Let us first introduce a specific notation. For an element $\mathbf{v}^* \in \operatorname{ran}(\mathbf{A}^*) \subset \mathcal{X}^*$ we can write $\mathbf{v}^* = \mathbf{A}^*\mathbf{o}^*$. We call then $\mathbf{o}^*$ in the following the *precursor* of $\mathbf{v}^*$ because of the intimate connection between the spaces $\mathcal{Y}^*$ and $\mathcal{X}^*$ via the adjoint operator $\mathbf{A}^*$.

The following optimality condition, stated in [1, Lemma 2.10], is of special importance.

**Lemma 2** (Optimality condition)**.** Let $\mathcal{X}$ be smooth and uniformly convex and $\mathbf{y} \in \operatorname{ran}(\mathbf{A})$.

(a) There exists the minimum-norm-solution $\mathbf{x}^\dagger \in \mathcal{X}$ of (2), i.e.

$$\|\mathbf{x}^\dagger\| = \min\big\{\|\mathbf{x}\| : \mathbf{A}\mathbf{x} = \mathbf{y},\ \mathbf{x} \in \mathcal{X}\big\},$$

and $J_p(\mathbf{x}^\dagger) \in \overline{\operatorname{ran}(\mathbf{A}^*)}$.

(b) If $\mathbf{x}^\dagger \in \mathcal{X}$ is the minimum-norm-solution of (2) and $\widetilde{\mathbf{x}} \in \mathcal{X}$ fulfills $J_p(\widetilde{\mathbf{x}}) \in \overline{\operatorname{ran}(\mathbf{A}^*)}$ and $\mathbf{x}^\dagger - \widetilde{\mathbf{x}} \in \operatorname{nul}(\mathbf{A})$, then $\widetilde{\mathbf{x}} = \mathbf{x}^\dagger$.

### *3.1. Sequential Subspace Optimization (SESOP)*

For convenience let us first recall the SESOP method as given in [4] for solving the ill-posed inverse problem $\mathbf{A}\mathbf{x} = \mathbf{y}$ without noise. The solution manifold of (2) is denoted by

$$\mathcal{M}_{\mathbf{A}\mathbf{x}=\mathbf{y}} := \big\{\mathbf{x} \in \mathcal{X} : \mathbf{A}\mathbf{x} = \mathbf{y}\big\}.$$

**Method 1** (SESOP)**.**

*(S1) Take $\mathbf{x}_0$ as initial value with $J_p(\mathbf{x}_0) \in \overline{ran\,(\mathbf{A}^*)}$, set $n := 0$, $\mathcal{U}_{-1} := \{0\}$ and repeat the following steps:*

*(S2) If $R_n := \|\mathbf{R}_n\| := \|\mathbf{A}\mathbf{x}_n - \mathbf{y}\| = 0$ then STOP else goto (S3).*

*(S3) Choose the search space $\mathcal{U}_n = \operatorname{Span}\{\mathbf{u}^*_{n,1}, \ldots, \mathbf{u}^*_{n,N_n}\} \subset ran\,(\mathbf{A}^*)$ with $N_n$ search directions $\mathbf{u}^*_{n,k} \in \mathcal{U}_n$, $k = 1, \ldots, N_n$ and with $N_n$ offsets $\alpha_{n,k} := \langle \mathbf{u}^*_{n,k}, \mathbf{z}\rangle$ for any $\mathbf{z} \in \mathcal{M}_{\mathbf{A}\mathbf{x}=\mathbf{y}} := \{\mathbf{x} \in \mathcal{X} : \mathbf{A}\mathbf{x} = \mathbf{y}\}$.*

*(S4) Compute the new iterate*

$$\mathbf{x}_{n+1} := J^*_{p^*}\Big(J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} \mu_{n,k}\mathbf{u}^*_{n,k}\Big) \tag{18}$$

*where $\mu_n = (\mu_{n,1}, \ldots, \mu_{n,N_n})$ is the solution of the $N_n$-dimensional optimization problem*

$$\min_{t\in\mathbb{R}^{N_n}} h_n(t)$$

*with*

$$h_n(t) := \frac{1}{p^*} \left\| J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} t_k \mathbf{u}^*_{n,k} \right\|^{p^*} + \sum_{k=1}^{N_n} t_k \alpha_{n,k} \tag{19}$$

$$\partial_j h_n(t) = -\left\langle \mathbf{u}^*_{j,k}, J^*_{p^*}\left( J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} t_k \mathbf{u}^*_{n,k} \right) \right\rangle + \alpha_{j,k} \quad \forall j = 1, \dots, N_n \tag{20}$$

*(S5) Set $n \leftarrow n+1$ and goto (S2).*

Note that $\mu_n$ is uniquely determined, since $h_n(t)$ (19) is strictly convex.

Convergence of the method essentially depends on the choice of the search space $\mathcal{U}_n$ and associated offsets $\alpha_n$ per iteration step $n$, see step (S3). We state a few common choices, taken from [4, 8], where $\mathbf{d}^*_n := \mathbf{A}^*\mathbf{R}^*_n = \mathbf{A}^* j_r\big(\mathbf{A}\mathbf{x}_n - \mathbf{y}\big)$ denotes the *Landweber direction* with precursor $\mathbf{R}^*_n = j_r(\mathbf{R}_n)$. With $j_r$ we define a single-valued selection of the set-valued duality mapping $J_r : Y \to 2^{\mathcal{Y}^*}$.

(a) *Expanding*: $\mathcal{U}^{\text{exp}}_n = \text{Span}\,\{\mathbf{d}^*_0, \dots, \mathbf{d}^*_n\}$, $\alpha^{\text{exp}}_{n,k} = \langle \mathbf{R}^*_{n,k}, \mathbf{y} \rangle$ with dimension $|\mathcal{U}_n| = n+1$

(b) *Truncated*: $\mathcal{U}^{\text{trunc}}_n = \text{Span}\,\{\mathbf{d}^*_{n-N_n+1}, \dots, \mathbf{d}^*_n\}$, $\alpha^{\text{trunc}}_{n,k} = \langle \mathbf{R}^*_{n,k}, \mathbf{y} \rangle$ with dimension $|\mathcal{U}_n| = N_n := N \wedge (n+1)$ for some fixed $N \in \mathbb{N}$

(c) *Nemirovsky I*: $\mathcal{U}^{\text{Nem1}}_n = \text{Span}\,\{\mathbf{d}^*_n, J_p(x_n) - J_p(x_0)\}$, $\alpha^{\text{Nem1}}_{n,k} = \langle \mathbf{v}^*_{n,k}, \mathbf{y} \rangle$ with dimension $|\mathcal{U}_n| = 2$, $\mathbf{A}^*\mathbf{v}_{n,1} = \mathbf{d}^*_n$, $\mathbf{A}^*\mathbf{v}_{n,2} = J_p(x_n) - J_p(x_0)$

(d) *Nemirovsky II*: $\mathcal{U}^{\text{Nem2}}_n = \text{Span}\,\{\mathbf{d}^*_n, J_p(x_n) - J_p(x_{n-1})\}$, $\alpha^{\text{Nem2}}_{n,k} = \langle \mathbf{v}^*_{n,k}, \mathbf{y} \rangle$ with dimension $|\mathcal{U}_n| = 2$, $\mathbf{A}^*\mathbf{v}_{n,1} = \mathbf{d}^*_n$, $\mathbf{A}^*\mathbf{v}_{n,2} = J_p(x_n) - J_p(x_{n-1})$

Note that the Nemirovsky directions of cases (c) and (d), that provide strong convergence, [8, Prop. 1], are not considered in this article. Furthermore, for finite-dimensional spaces $\mathcal{X}$ and $\mathcal{Y}$ weak and strong convergence coincide [19, Thm 4.3]. In the cases (a) or (b) the hyperplane offsets can simply be calculated by

$$\alpha_{n,k} := \langle \mathbf{R}^*_{n,k}, \mathbf{y} \rangle. \tag{21}$$

### *3.2. Orthogonality in Banach spaces*

Incorporating multiple search directions into the regularization process significantly improves convergence speed as was demonstrated by experiments [4, Sect. 5]. However, the search directions used there are not related to each other. In the following we would like to maximize *distinctiveness* of the search directions $\mathbf{u}^*_{n,k}$ in the truncated search space $\mathcal{U}_n$, i. e. we *orthogonalize* them with respect to previous directions contained in $\mathcal{U}_n$ and to some extent to older search directions that are not contained in $\mathcal{U}_n$.

The concept of *orthogonality* has already been generalized to Banach spaces in [23, 24, 25]. Among these we see the one of [25], that relies on [24], as appropriate because of its relation to the metric projection which is shown later on.

**Definition 8** (j-orthogonality [12, Def. 2.3])**.** We say that an element $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is a Banach space, is *(j-)orthogonal* to $\mathbf{y} \in \mathcal{X}$ if

$$\|\mathbf{x}\| \leq \|\mathbf{x} + t\mathbf{y}\| \qquad \text{for all } t \in \mathbb{R}. \tag{22}$$

We note that in smooth Banach spaces j-orthogonality is equivalent to the generalized orthogonality definition given by [12, p. 335], which is the case in our setting of orthogonalizing gradient directions in a uniform smooth dual space $\mathcal{X}^*$.

**Definition 9** (Generalized Orthogonality,[12, Def. 2.2])**.** An element $\mathbf{x} \in \mathcal{X}$ in a Banach space $\mathcal{X}$ is said to be *(g-)orthogonal* to $\mathbf{y} \in \mathcal{X}$ if $\langle J_p(\mathbf{x}), \mathbf{y} \rangle = 0$.

There exists a relation between the metric projection and g-orthogonality that is of great importance for our further considerations.

**Proposition 4** (Metric projection and g-orthogonality, [12, Prop. 2.10])**.** Let $\mathcal{X}$ be a Banach space. If $\mathbf{x} \in \mathcal{X}$ and $\mathcal{M} \subset \mathcal{X}$ is a closed subspace, then $\widetilde{\mathbf{x}} = P_{\mathcal{M}}(\mathbf{x})$, if and only if $\widetilde{\mathbf{x}} \in \mathcal{M}$ and

$$\langle J_p(\mathbf{x} - \widetilde{\mathbf{x}}), \mathbf{v} \rangle = 0 \qquad \text{for all } \mathbf{v} \in \mathcal{M}, \tag{23}$$

i.e. the error $\mathbf{x} - \widetilde{\mathbf{x}}$ is g-orthogonal to $\mathcal{M}$.

Proposition 4 implies that in Banach spaces an element $\mathbf{x}$ can be made (g-) orthogonal to another element $\mathbf{y}$ or a subspace $\mathcal{M}$ by subtracting its metric projection. We like to stress that by using the concept of g-orthogonality we are able to orthogonalize a new search direction with respect to given directions in the search space, but this does not imply pairwise orthogonality of *all* search directions to each other. The reason is that Definition 9 is not symmetric, since in general $0 = \langle J_p(\mathbf{x}), \mathbf{y} \rangle \neq \langle J_p(\mathbf{y}), \mathbf{x} \rangle$. We always denote by $\mathcal{V}_n^{\mathrm{trunc}}$ the search space consisting of directions that are g-orthogonalized in contrast to $\mathcal{U}_n^{\mathrm{trunc}}$, c. f. section 3.1, which consists of $N_n$ non-orthogonalized directions.

### 3.3. Orthogonalization of the search directions

Based on the definition of g-orthogonality, see Definition 9, we outline the construction of a search space, which is similar to $\mathcal{U}_n^{\mathrm{trunc}}$, but consists of orthogonalized directions. To this end, we investigate search directions which are deduced from the Landweber direction $\mathbf{d}_n^* = \mathbf{A}^* j_r\big(\mathbf{A}\mathbf{x}_n - \mathbf{y}\big)$ to obtain the orthogonalized search space $\mathcal{V}_n^{\mathrm{trunc}} := \{\mathbf{v}_{n,k}^*\}_{k=1}^{N_n}$. This is done by a procedure which is similar to the Gram-Schmidt orthogonalization process:

$$\begin{aligned} \mathbf{v}_{n,k-1}^* &:= \mathbf{v}_{n-1,k-(N_n-N_{n-1})}^* \quad \text{for } k = 2, \dots, N_n, \quad \text{if } N_n \geq 2 && (24) \\ \mathbf{v}_{n,N_n}^* &:= \mathbf{d}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{v}_{n-1,k}^* && (25) \\ &= \mathbf{d}_n^* - P_{\mathcal{V}_{n-1}^{\mathrm{trunc}}}\left(\mathbf{d}_n^*\right). && (26) \end{aligned}$$

Note that this requires $N_n - N_{n-1} \in \{0, 1\}$, i. e. the desired number of search directions remains constant or increases by one in each step $n$. If just a single search direction is used, $N_n = 1$, then (24) is not executed at all. Here, $s_{n,k}$ is the orthogonalization coefficient obtained from the metric projection of the Landweber direction $\mathbf{d}_n^*$ onto the search space $\mathcal{V}_{n-1}^{\mathrm{trunc}} = \{\mathbf{v}_{n-1,k}^*\}_{k=1}^{N_{n-1}}$. It is defined as

$$s_n = \big(s_{n,1}, \dots, s_{n,N_{n-1}}\big) := \arg \min_{s \in \mathbb{R}^{N_{n-1}}} g_n(s), \tag{27}$$

where

$$g_n(s) := \frac{1}{p^*} \left\| \mathbf{d}_n^* - \sum_{i=1}^{N_{n-1}} s_i \mathbf{v}_{n-1,i}^* \right\|_{\mathcal{X}^*}^{p^*}. \tag{28}$$

In Hilbert spaces this coincides with the familiar Gram-Schmidt procedure, see also [26, p. 218].

We show that all search directions $\mathbf{v}_{n,j}^*$, $j = 1, \ldots, N_n$, are g-orthogonal in the sense of Definition 9. This is the counterpart of the conjugacy property in Banach spaces when the CG method is applied to the normal equation, c. f. [6, p. 102].

**Corollary 1** (Orthogonal search directions)**.** We have

$$\langle \mathbf{v}_{n,j}^*, J_{p^*}^*(\mathbf{v}_{n,k}^*) \rangle = 0 \quad \forall 1 \le j < k \le N_n. \tag{29}$$

*Proof.* By means of Proposition 4 this follows directly from the optimality condition on $g_n(s)$ for (27),

$$\begin{aligned} 0 = \partial_j g_n(s_n) &= -\left\langle \mathbf{v}_{n-1,j}^*, J_{p^*}^* \left( \mathbf{d}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{v}_{n-1,k}^* \right) \right\rangle \\ &= -\langle \mathbf{v}_{n-1,j}^*, J_{p^*}^* \left( \mathbf{v}_{n,N_n}^* \right) \rangle \end{aligned} \tag{30}$$

for all $j = 1, \ldots, N_{n-1}$. As this holds for all $n$ and by the successive construction (24) of the search spaces $\mathcal{V}_n^{\text{trunc}}$, we have by induction that all search directions $\mathbf{v}_{n,k}^*$ in $\mathcal{V}_n^{\text{trunc}}$ with $1 \le k \le N_n$ are g-orthogonal with respect to $\mathbf{v}_{n,j}^*$ for $1 \le j < k$. □

Note that this orthogonality becomes symmetric, if $\mathcal{X}$ is a Hilbert space and $p = 2$, since $J_2 = \text{id}$ is the identity.

We point out that the summation in (25) is done over all directions in the search space $\mathcal{V}_{n-1}^{\text{trunc}}$, including $\mathbf{v}_{n-1,1}^*$, which is not contained in $\mathcal{V}_n^{\text{trunc}}$, see (24). Otherwise, the orthogonalization would not change the search space but only modify its spanning vectors. Hence, we obviously have that in general

$$\mathcal{U}_n^{\text{trunc}} \not\subset \mathcal{V}_n^{\text{trunc}} \qquad \text{for } n \ge 1.$$

Using the expanding search space $\mathcal{U}_n^{\text{exp}}$ on the other hand, see case (a) in Section 3.1, the above orthogonalization process of the new search direction would *not* change the iteration, i. e. for every $n$ we have $\mathcal{U}_n^{\text{exp}} = \mathcal{V}_n^{\text{exp}}$.

*Example* 1. In order to highlight the (notational) equivalence with the family of CG methods in Hilbert spaces, let us consider a single search direction, i. e. $\mathcal{V}_n^{\text{trunc}} = \operatorname{Span}\{\mathbf{v}_n^*\}$, per iteration step $n$. We then obtain

$$\mathbf{v}_n^* = \mathbf{d}_n^* - s_n \mathbf{v}_{n-1}^*. \tag{31}$$

So, we have the Landweber direction $\mathbf{d}_n^*$ which is modified by the last search direction $\mathbf{v}_{n-1}^*$ scaled by the orthogonalization coefficient $s_n$, i. e. $\mathbf{g_n} = \mathbf{d_n} - s_n \mathbf{g_{n-1}}$ where $\mathbf{g_n}$ is the current search direction and $\mathbf{d_n}$ is the current gradient direction in the usual notation of CG methods in Hilbert spaces, c. f. [6, Chap. 5] and also [27].

With each search direction an offset $\alpha_{n,N_n}$ is required that is connected to the solution manifold $\mathcal{M}_{\mathbf{Ax=y}}$, c. f. step (S3) in Method 1. Calculating the new hyperplane offsets $\beta_{n,k}$ to each orthogonalized search direction $\mathbf{v}_{n,k}^* \in \mathcal{V}_n^{\text{trunc}}$ is then done as

$$\beta_{n,N_n} := \langle \mathbf{o}_{n,N_n}^*, \mathbf{y}\rangle = \langle \mathbf{R}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{o}_{n-1,k}^*, \mathbf{y}\rangle = \alpha_{n,N_n} - \sum_{k=1}^{N_{n-1}} s_{n,k} \beta_{n-1,k}, \tag{32}$$

where we used

$$\begin{aligned} \mathbf{v}_{n,N_n}^* &= \mathbf{d}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{v}_{n-1,k}^* = \mathbf{d}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{A}^* \mathbf{o}_{n-1,k}^* \\ &= \mathbf{A}^* \left( \mathbf{R}_n^* - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{o}_{n-1,k}^* \right) =: \mathbf{A}^* \mathbf{o}_{n,N_n}^*, \end{aligned} \tag{33}$$

with the precursor $\mathbf{R}_n^* = j_r(\mathbf{R}_n)$ of the Landweber direction, $\mathbf{d}_n^* = \mathbf{A}^*\mathbf{R}_n^*$ and precursors $\mathbf{o}_{n-1,k}^*$ of $\mathbf{v}_{n-1,k}^*$.

The necessary orthogonalization coefficients $s_{n,k}$ are calculated by minimizing (28) with derivative (30) using standard techniques. Note that the proof of Lemma 4 yields a good initial guess for the according line search problem. Once the coefficients $s_{n,k}$ in (25) are computed, we can easily evaluate (32), knowing all other offsets $\beta_{n-1,k}$, $k = 1, \dots, N_{n-1}$, from previous iterations.

For completeness, we provide the full algorithm using orthogonalized search directions and a truncated search space $\mathcal{V}_n^{\text{trunc}}$ in Method 2.

**Method 2** (orthogonalized SESOP with $N_n$ search directions)**.**

*(S1) Take* $\mathbf{x}_0$ *as initial value with* $J_p(\mathbf{x}_0) \in \overline{ran\,(\mathbf{A}^*)}$*, set* $n := 0$*,* $N_0 := 1$*,* $\mathcal{V}_{-1}^{trunc} := \{0\}$ *and repeat the following steps:*

*(S2) If* $R_n := \|\mathbf{R}_n\| := \|\mathbf{A}\mathbf{x}_n - \mathbf{y}\| = 0$ *then STOP else goto (S3).*

*(S3) Orthogonalize the Landweber direction* $\mathbf{d}_n^* = \mathbf{A}^* j_r\big(\mathbf{A}\mathbf{x}_n - \mathbf{y}\big)$ *using* $\mathcal{V}_{n-1}^{trunc}$*, see (25), to obtain* $\mathbf{v}_{n,N_n}^*$*.*

*(S4) Update the search space* $\mathcal{V}_n^{trunc} = \operatorname{Span}\{\mathbf{v}_{n,1}^*, \dots, \mathbf{v}_{n,N_n}^*\}$ *by using the last* $N_n - 1$ *directions from* $\mathcal{V}_{n-1}^{trunc}$*, see (24), and adding the orthogonalized Landweber direction* $\mathbf{v}_{n,N_n}^*$*.*

*(S5) In a similar manner, update the hyperplane offsets to the orthogonalized search directions* $\beta_{n,k}$ *using (32) for the orthogonalized Landweber direction* $\mathbf{v}_{n,N_n}^*$*.*

*(S6) Compute the new iterate*

$$\mathbf{x}_{n+1} := J_{p^*}^*\Big(J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} \mu_{n,k}\mathbf{v}_{n,k}^*\Big) \tag{34}$$

*where* $\mu_n = (\mu_{n,1}, \dots, \mu_{n,N_n})$ *is the solution of the* $N_n$*-dimensional optimization problem*

$$\min_{t\in\mathbb{R}^{N_n}} h_n(t)$$

*with*

$$h_n(t) := \frac{1}{p^*}\left\|J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} t_k\mathbf{v}_{n,k}^*\right\|^{p^*} + \sum_{k=1}^{N_n} t_k\beta_{n,k} \tag{35}$$

$$\partial_j h_n(t) = -\left\langle \mathbf{v}_{j,k}^*, J_{p^*}^*\left(J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} t_k\mathbf{v}_{n,k}^*\right)\right\rangle + \beta_{j,k} \quad \forall j = 1, \dots, N_n \tag{36}$$

*(S7) Set $n \leftarrow n+1$ and goto (S2).*

We note that because only a single new direction can be added to the search space $\mathcal{V}_n^{\text{trunc}}$ per step, $N_n$ may increase at most by one per step. Therefore, if a fixed number of four search directions are to be used during the optimization, then we employ the sequence $N_n = \{1, 1, 2, 3, 4, 4, \dots, 4\}_{n=0}$.

### *3.4. Proof of convergence*

In this section we prove that Method 1 equipped with the orthogonalized truncated search space $\mathcal{V}_n^{\text{trunc}}$ still converges weakly to a solution of $\mathbf{Ax} = \mathbf{y}$.

We note that $R_n = 0$ implies that $\mathbf{Ax}_n = \mathbf{y}$ and we are done. So, this yields a good stopping criterion. If $R_n \neq 0$, then we also have for the Landweber direction $\mathbf{d}_n^* \neq 0$.

**Corollary 2.** If $R_n \neq 0$, then we have $\mathbf{d}_n^* \neq 0$.

*Proof.* Assume the contrary, $\mathbf{d}_n^* = 0$, and let $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax}=\mathbf{y}}$. Then we get

$$0 = \langle \mathbf{d}_n^*, \mathbf{x}_n - \mathbf{z} \rangle = \langle \mathbf{R}_n^*, \mathbf{Ax}_n - \mathbf{Az} \rangle = \|\mathbf{R}_n\|^r = R_n^r,$$

which is a contradiction. Note that by the very same argument we also have $\mathbf{R}_n^* \neq 0$. □

Next we prove something similar to the "expanding subspace" property of CG methods, c. f. [6, Theorem 5.2].

**Corollary 3.** Let $\mathcal{V}_n^{\text{trunc}} := \text{Span}\,\{\mathbf{v}_{n,1}^*, \dots, \mathbf{v}_{n,N_n}^*\}$ be a g-orthogonalized search space with precursors $\{\mathbf{o}_{n,i}^*\}_{i=1}^{N_n}$. Then, the following assertions hold true.

(a) We have that

$$\begin{aligned} &\langle \mathbf{o}_{n-1,1}^*, \mathbf{R}_{n-N_n} \rangle = 0 \\ &\dots \\ &\langle \mathbf{o}_{n-1,1}^*, \mathbf{R}_n \rangle = 0, \dots, \langle \mathbf{o}_{n-1,N_{n-1}}^*, \mathbf{R}_n \rangle = 0. \end{aligned} \tag{37}$$

Note that this extends to $\{\mathbf{o}_{n,1}^*, \dots, \mathbf{o}_{n,N_n-1}^*\}$ by construction of $\mathcal{V}_n^{\text{trunc}}$.

(b) If $R_n \neq 0$, then we also have $\mathbf{v}^*_{n,N_n} \neq 0$.

(c) Each of the sets $\{\mathbf{o}^*_{n,1}, \dots, \mathbf{o}^*_{n,N_n}\}$ and $\{\mathbf{v}^*_{n,1}, \dots, \mathbf{v}^*_{n,N_n}\}$ is linearly independent.

*Proof.* We first prove (a). Assume $R_n \neq 0$, i. e. $\mathbf{R}_n \neq 0$. The optimality condition of the step size functional $h_{n-1}(\mu)$, see (19), at $n-1$ for $j = 1, \dots, N_{n-1}$ is given by

$$\begin{aligned}
0 &= \partial_j h_{n-1}(\mu_{n-1}) \\
&= -\langle \mathbf{v}^*_{n-1,j}, J^*_{p^*}\Big(J_p(\mathbf{x}_{n-1}) - \sum_{k=1}^{N_{n-1}} \mu_{n-1,k}\mathbf{v}^*_{n-1,k}\Big)\rangle + \beta_{n-1,j} \\
&= -\langle \mathbf{o}^*_{n-1,j}, \mathbf{A}\mathbf{x}_n\rangle + \langle \mathbf{o}^*_{n-1,j}, \mathbf{y}\rangle \\
&= -\langle \mathbf{o}^*_{n-1,j}, \mathbf{R}_n\rangle,
\end{aligned}$$

where we used (18) and (32). The statement then follows by (24) and stepping back until $n - N_n$ if we take into account that

$$\mathcal{V}_n^{\text{trunc}} = \{\mathbf{v}^*_{n-N_n-1, N_{n-N_n}-1}, \dots, \mathbf{v}^*_{n,N_n}\}.$$

We continue with (b). By Corollary 2 we have $\mathbf{d}^*_n \neq 0$ and, thus, by (25) we have to show that $\mathbf{v}^*_{n,N_n}$ is not contained in $\mathcal{V}_{n-1}$. To this end, let $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$ and hence $\mathbf{x}_n - \mathbf{z} \neq 0$. Furthermore, let $\lambda_1, \dots, \lambda_{N_{n-1}}, \sigma \in \mathbb{R}$ be given with

$$\sum_{k=1}^{N_{n-1}} \lambda_k \mathbf{v}^*_{n-1,k} + \sigma \mathbf{d}^*_n = 0.$$

Then, by using (a),

$$\begin{aligned}
0 &= \sum_{k=1}^{N_{n-1}} \lambda_k \langle \mathbf{v}^*_{n-1,k}, \mathbf{x}_n - \mathbf{z}\rangle + \sigma\langle \mathbf{d}^*_n, \mathbf{x}_n - \mathbf{z}\rangle \\
&= \sum_{k=1}^{N_{n-1}} \lambda_k \langle \mathbf{o}^*_{n-1,k}, \mathbf{R}_n\rangle + \sigma\langle j_r(\mathbf{R}_n), \mathbf{R}_n\rangle \\
&= \sigma R_n^r,
\end{aligned}$$

we get $\sigma = 0$, which proves the statement by contradiction.

As for (c) it suffices to show that $\{\mathbf{v}_{n,1}^*, \dots, \mathbf{v}_{n,N_n}^*\}$ are linearly independent. Assume again $R_n \neq 0$ and let $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$ and $\lambda_1, \dots, \lambda_{N_n} \in \mathbb{R}$ be given with

$$\sum_{k=1}^{N_n} \lambda_k \mathbf{v}_{n,k}^* = 0.$$

Applying (25) and (a) to $\mathbf{v}_{n,N_n}^*$, we deduce

$$\begin{aligned} 0 &= \sum_{k=1}^{N_n} \lambda_k \langle \mathbf{v}_{n,k}^*, \mathbf{x}_n - \mathbf{z} \rangle = \sum_{k=1}^{N_n - 1} \lambda_k \langle \mathbf{o}_{n,k}^*, \mathbf{R}_n \rangle + \lambda_{N_n} \langle \mathbf{o}_{n,N_n}^*, \mathbf{R}_n \rangle \\ &= \lambda_{N_n} \langle \mathbf{o}_{n,N_n}^*, \mathbf{R}_n \rangle = \lambda_{N_n} \langle \mathbf{R}_n^*, \mathbf{R}_n \rangle - \sum_{k=1}^{N_n} s_{n,k} \langle \mathbf{o}_{n-1,k}^*, \mathbf{R}_n \rangle \\ &= \lambda_{N_n} R_n^r, \end{aligned}$$

what implies $\lambda_{N_n} = 0$. We continue with

$$0 = \sum_{k=0}^{N_n - 1} \lambda_k \langle \mathbf{v}_{n,k}^*, \mathbf{x}_{n-1} - \mathbf{z} \rangle = \ldots = \lambda_{N_n - 1} R_{n-1}^r,$$

and by induction we get $\lambda_k = 0$ for all $k = 1, \dots, N_n$. Hence, also the search directions $\mathbf{v}_{n,k}^*$ are linearly independent. □

We continue by showing that our solution manifold is contained in the intersection of certain hyperplanes and that the iterates and search directions obtained via the update formulas (18) and (25) can be interpreted as Bregman projections on intersections of hyperplanes. Though most of this is not needed in the convergence proof, it is very illustrative for the general iteration process. For $\mathbf{x}^* \in \mathcal{X}^*$ and $\beta \in \mathbb{R}$ we denote by

$$\mathcal{H}(\mathbf{x}^*, \beta) := \left\{ \mathbf{x} \in \mathcal{X} : \langle \mathbf{x}^*, \mathbf{x} \rangle = \beta \right\}$$

the *hyperplane* defined by $\mathbf{x}^*$ and offset $\beta$.

**Lemma 3** (Intersection of hyperplanes)**.**

(a) For the solution manifold $\mathcal{M}_{\mathbf{Ax=y}}$ we have

$$\mathcal{M}_{\mathbf{Ax=y}} \subset \mathcal{H}_n := \bigcap_{k=1}^{N_n} \mathcal{H}(\mathbf{v}_{n,k}^*, \beta_{n,k}). \tag{38}$$

(b) For all $n$ it holds $J_p(\mathbf{x}_n) - J_p(\mathbf{x}_0) \in \bigcup_n \mathcal{V}_n \subset \overline{\operatorname{ran}(\mathbf{A}^*)}$.

(c) The iterate $\mathbf{x}_{n+1}$ is the Bregman projection of the current iterate $\mathbf{x}_n$ onto $\mathcal{H}_n$,

$$\mathbf{x}_{n+1} = \Pi^p_{\mathcal{H}_n}(\mathbf{x}_n) \tag{39}$$

and furthermore

$$J_p(\mathbf{x}_{n+1}) = \Pi^p_{J_p(\mathbf{x}_n)+\mathcal{V}_n^{\mathrm{trunc}}}\left(J_p(\mathbf{z})\right) \qquad \text{for all } \mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}. \tag{40}$$

(d) The search direction $\mathbf{v}^*_{n,N_n}$ is the Bregman projection of the Landweber direction $\mathbf{d}^*_n$,

$$J^*_{p^*}(\mathbf{v}^*_{n,N_n}) = \Pi^p_{(\mathcal{V}_{n-1}^{\mathrm{trunc}})^\perp}\left(J^*_{p^*}(\mathbf{d}^*_n)\right), \tag{41}$$

where $\mathcal{V}^\perp$ denotes the annihilator of the space $\mathcal{V}$.

*Proof.* To prove part (a) we have to show $\langle \mathbf{v}^*_{n,k}, \mathbf{z}\rangle = \beta_{n,k}$ for any $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$, what follows directly from the definition of the offsets, (32).

Part (b) follows from

$$\begin{aligned} J_p(\mathbf{x}_n) - J_p(\mathbf{x}_0) &= J_p(\mathbf{x}_n) - J_p(\mathbf{x}_{n-1}) + J_p(\mathbf{x}_{n-1}) - \ldots + J_p(\mathbf{x}_1) - J_p(\mathbf{x}_0), \\ &= \sum_{j=1}^{N_n} \mu_n \mathbf{v}^*_{n,j} + \ldots + \sum_{j=1}^{N_1} \mu_1 \mathbf{v}^*_{1,j}, \end{aligned}$$

c. f. (18).

Part (d) is proven by using the definition (25) of $\mathbf{v}^*_{n,N_n}$, relations between metric and Bregman projections, see Prop. 3(a) and [4, Prop. 3.6 d)] as well as known equivalencies for Bregman projections, see [4, Prop. 3.7 b)]. Putting these ingredients together we obtain

$$\begin{aligned} J^*_{p^*}(\mathbf{v}^*_{n,N_n}) &= J^*_{p^*}\left(\mathbf{d}^*_n - P_{\mathcal{V}_{n-1}^{\mathrm{trunc}}}(\mathbf{d}^*_n)\right) = J^*_{p^*}\left(-\Pi_{\mathcal{V}_{n-1}^{\mathrm{trunc}}-\mathbf{d}^*_n}(0)\right) \\ &= J^*_{p^*}\left(\Pi_{\mathbf{d}^*_n-\mathcal{V}_{n-1}^{\mathrm{trunc}}}(0)\right) = J^*_{p^*}\left(\Pi_{\mathbf{d}^*_n+\mathcal{V}_{n-1}^{\mathrm{trunc}}}(0)\right) \\ &= \Pi_{(\mathcal{V}_{n-1}^{\mathrm{trunc}})^\perp}\left(J^*_{p^*}(\mathbf{d}^*_n)\right). \end{aligned}$$

Part (c) follows in the same way as in the proof of [4, Prop. 4.1]. □

As last ingredient for our convergence proof we need to show that $\mathbf{v}^*_{n,N_n}$ is still a descent direction, where we use the same geometrical arguments as in the generalized Landweber convergence proof, see [1]. We emphasize that this is not a straight-forward consequence, since we have $\mathcal{U}_n^{\mathrm{trunc}} \not\subset \mathcal{V}_n^{\mathrm{trunc}}$, i. e. $\mathbf{d}^*_n$ is in general not contained in $\mathcal{V}_n^{\mathrm{trunc}}$, c. f. (26).

**Lemma 4** (Descent direction property)**.** Any $\mathbf{v}^*_{n,N_n}$ resulting from (25) is always a descent direction, i. e. there is a $\mu_n \in \mathbb{R}^{N_n}$ and $S_n > 0$ with

$$h_n(\mu_n) \leq h_n(0) - S_n \tag{42}$$

For any $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$ this is equivalent to

$$\Delta_p(\mathbf{x}_{n+1}, \mathbf{z}) \leq \Delta_p(\mathbf{x}_n, \mathbf{z}) - S_n. \tag{43}$$

*Proof.* We assume $R_n \neq 0$ and $\mathbf{x}_n \neq 0$. We set

$$\widetilde{\mu}_n := (0, \ldots, 0, \nu_n) \quad \text{with } \nu_n := \frac{\tau_n \left\|\mathbf{x}_n\right\|_{\mathcal{X}}^{p-1}}{\left\|\mathbf{v}^*_{n,N_n}\right\|_{\mathcal{X}^*}} \tag{44}$$

where $\tau_n \in (0, 1]$ is to be chosen satisfying

$$\frac{\rho_{\mathcal{X}^*}(\tau_n)}{\tau_n} = \rho_{\mathcal{X}^*}(1) \wedge \left( \frac{\gamma}{2^{p^*} G_p} \frac{R_n^r}{\left\|\mathbf{x}_n\right\|_{\mathcal{X}} \left\|\mathbf{v}^*_n\right\|_{\mathcal{X}^*}} \right).$$

for $\gamma \in (0, 1)$, c. f. Theorem 2. Let $\mu_n = \arg\min_{t \in \mathbb{R}^n} h_n(t)$, then we estimate by means of the Xu-Roach inequality (8)

$$\begin{aligned} h_n(\mu_n) &\leq h_n(\widetilde{\mu}_n) \\ &\leq \tfrac{1}{p^*} \left\|\mathbf{x}_n\right\|_{\mathcal{X}}^p - \nu_n \langle \mathbf{v}^*_{n,N_n}, \mathbf{x}_n \rangle + \frac{1}{p^*} \widetilde{\sigma} \left( J_p(\mathbf{x}_n), \nu_n \mathbf{v}^*_{n,N_n} \right) + \langle \mathbf{o}^*_{n,N_n}, \mathbf{y} \rangle \\ &= \tfrac{1}{p^*} \left\|\mathbf{x}_n\right\|_{\mathcal{X}}^p - \nu_n \langle \mathbf{R}^*_n - \sum_{k=1}^{N_{n-1}} s_{n,k} \mathbf{o}^*_{n-1,k}, \mathbf{R}_n \rangle + \frac{1}{p^*} \widetilde{\sigma} \left( J_p(\mathbf{x}_n), \nu_n \mathbf{v}^*_{n,N_n} \right) \\ &= \tfrac{1}{p^*} \left\|\mathbf{x}_n\right\|_{\mathcal{X}}^p - \nu_n R_n^r + \frac{1}{p^*} \widetilde{\sigma} \left( J_p(\mathbf{x}_n), \nu_n \mathbf{v}^*_{n,N_n} \right) \\ &= \tfrac{1}{p^*} \left\|\mathbf{x}_n\right\|_{\mathcal{X}}^p - \tau_n \left\|\mathbf{x}_n\right\|^{p-1} \frac{R_n^r}{\left\|\mathbf{v}^*_{n,N_n}\right\|_{\mathcal{X}^*}} + \frac{1}{p^*} \widetilde{\sigma} \left( J_p(\mathbf{x}_n), \nu_n \mathbf{v}^*_{n,N_n} \right), \end{aligned}$$

where we have used the orthogonality properties stated in Corollary 3(a). As the metric projection is non-expanding, c. f. (26), we have that

$$\left\|\mathbf{v}^*_{n,N_n}\right\|_{\mathcal{X}^*} \le \left\|\mathbf{A}^* j_r(\mathbf{A}\mathbf{x}_n - \mathbf{y})\right\|_{\mathcal{X}^*} \le \|\mathbf{A}\| \, R_n^{r-1}.$$

Lemma 1 allows us to bound the last summand using the requirement on $\tau_n$,

$$\begin{aligned}
\tfrac{1}{p^*}\widetilde{\sigma}_{p^*}\left(J_p(\mathbf{x}_n), \nu_n \mathbf{v}^*_{n,N_n}\right) &\le 2^{p^*} G_p \left\|\mathbf{x}_n\right\|^p_{\mathcal{X}} \rho_{\mathcal{X}^*}(\tau_n) \\
&\le \tau_n 2^{p^*} G_p \left\|\mathbf{x}_n\right\|^p_{\mathcal{X}} \frac{\rho_{\mathcal{X}^*}(\tau_n)}{\tau_n} \\
&\le \tau_n 2^{p^*} G_p \left\|\mathbf{x}_n\right\|^p_{\mathcal{X}} \frac{\gamma}{2^{p^*} G_p} \frac{R_n^r}{\left\|\mathbf{x}_n\right\|_{\mathcal{X}} \left\|\mathbf{v}^*_{n,N_n}\right\|_{\mathcal{X}^*}} \\
&= \gamma \tau_n \left\|\mathbf{x}_n\right\|^{p-1}_{\mathcal{X}} \frac{R_n^r}{\left\|\mathbf{v}^*_{n,N_n}\right\|_{\mathcal{X}^*}}.
\end{aligned}$$

Combining all that we obtain

$$\begin{aligned}
h_n(\mu_n) &\le \tfrac{1}{p^*}\left\|\mathbf{x}_n\right\|^p_{\mathcal{X}} - (1-\gamma)\tau_n \left\|\mathbf{x}_n\right\|^{p-1}_{\mathcal{X}} \frac{R_n^r}{\left\|\mathbf{v}^*_{n,N_n}\right\|} \\
&\le \tfrac{1}{p^*}\left\|\mathbf{x}_n\right\|^p_{\mathcal{X}} - \frac{(1-\gamma)}{\|\mathbf{A}\|}\tau_n \left\|\mathbf{x}_n\right\|^{p-1}_{\mathcal{X}} R_n. \qquad (45)
\end{aligned}$$

We are done since all factors in $S_n := \frac{(1-\gamma)}{\|\mathbf{A}\|}\tau_n \left\|\mathbf{x}_n\right\|^{p-1}_{\mathcal{X}} R_n$ are positive and if we take $h_n(0) = \frac{1}{p^*}\left\|\mathbf{x}_n\right\|^p_{\mathcal{X}}$ into account.

It remains to show (43). To this end we use (14) and (18). For any $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax}=\mathbf{y}}$ we then have

$$\begin{aligned}
\Delta_p(\mathbf{x}_{n+1}(\mu), \mathbf{z}) &= \tfrac{1}{p^*}\left\|\mathbf{x}_{n+1}(\mu)\right\|^p_{\mathcal{X}} - \langle J_p(\mathbf{x}_n) - \sum_{k=1}^{N_n} \mu_k \mathbf{v}^*_{n,k}, \mathbf{z}\rangle + \tfrac{1}{p}\left\|\mathbf{z}\right\|^p_{\mathcal{X}} \\
&= \tfrac{1}{p^*}\left\|\mathbf{x}_{n+1}(\mu)\right\|^p_{\mathcal{X}} - \langle J_p(\mathbf{x}_n), \mathbf{z}\rangle + \sum_{k=1}^{N_n} \mu_k \beta_{n,k} + \tfrac{1}{p}\left\|\mathbf{z}\right\|^p_{\mathcal{X}} \\
&= h_n(\mu) - \langle J_p(\mathbf{x}_n), \mathbf{z}\rangle + \tfrac{1}{p}\left\|\mathbf{z}\right\|^p_{\mathcal{X}}.
\end{aligned}$$

As the last two terms are constant with respect to $\mu$, they cancel out when considering the difference $h_n(0) - h_n(\mu)$. This finally yields (43). □

Our investigations are subsumed in the main result of this section.

**Theorem 3** (Weak convergence for $\mathcal{V}_n^{\text{trunc}}$)**.** Let $\mathcal{X}$ be a uniformly convex and smooth Banach space with sequentially weak-to-weak continuous duality mapping $J_p$ and $\mathcal{Y}$ be an arbitrary Banach space. For $1 \leq N_n \leq n$ and search space $\mathcal{V}_n^{\text{trunc}}$ given by (24) and (25), Method 1 either stops after a finite number $n \in \mathbb{N}$ of iterations (in case $R_n = 0$) with $\mathbf{x}_n = \widehat{\mathbf{x}} = \Pi^p_{\mathcal{M}_{\mathbf{Ax=y}}}(\mathbf{x}_0)$ being the Bregman projection of $\mathbf{x}_0$ onto the solution manifold $\mathcal{M}_{\mathbf{Ax=y}}$ or the sequence of the iterates $\{\mathbf{x}_n\}_n$ converges weakly to $\widehat{\mathbf{x}}$.

*Proof.* In case $R_n = 0$ for some $n \in \mathbb{N}$ we have $\mathbf{x}_n \in \mathcal{M}_{\mathbf{Ax=y}}$ and we are done by [4, Proposition 3.7 b)] together with the optimality condition in Lemma 2(b).

Let us therefore assume $R_n \neq 0$ for all $n$. Lemma 4 ensures that $\{\Delta_p(\mathbf{x}_n, \mathbf{z})\}_n$ for all $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$ is strictly decreasing. Thus, $\{\Delta_p(\mathbf{x}_n, \mathbf{z})\}_n$ is bounded from above by $\{\Delta_p(\mathbf{x}_0, \mathbf{z})\}_n$. Then, Proposition 2 (b) states that $\{\mathbf{x}_n\}_n$ is bounded. As we require $\mathcal{X}$ to be uniformly convex and hence reflexive by the Milman-Pettis theorem, [13, Sect. II.2, Thm. 2.9], every subsequence of $\{\mathbf{x}_n\}_n$ has a subsequence $\{\mathbf{x}_{n_k}\}_k$ that converges weakly to some $\widehat{\mathbf{x}} \in \mathcal{X}$, see [19, Chap. 8,Thm 4.2]. The proof of $\{R_{n_k}\}_k$ being a null sequence follows in exactly the same way as in [1, p. 320]. As a consequence we even have $\widehat{\mathbf{x}} \in \mathcal{M}_{\mathbf{Ax=y}}$ as $R_{n_k} = \|\mathbf{A}\mathbf{x}_{n_k} - \mathbf{y}\|_{\mathcal{Y}} \to 0$ with $k \to \infty$, i. e. $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$. As $\overline{\operatorname{ran}(\mathbf{A}^*)}$ is convex and norm-closed, it is also weakly closed, see [28, Chap. 5, Thm. 3.13]. This together with Lemma 3 (b) implies $J_p(\widehat{\mathbf{x}}) - J_p(\mathbf{x}_0) \in \overline{\operatorname{ran}(\mathbf{A}^*)}$ since $J_p$ is sequentially weak-to-weak continuous by assumption. Applying the optimality condition Lemma 2(b) with the requirement $J_p(\mathbf{x}_0) \in \overline{\operatorname{ran}(\mathbf{A}^*)}$ and taking into account that $\mathbf{z} + \operatorname{nul}(\mathbf{A}) = \mathcal{M}_{\mathbf{Ax=y}}$ for all $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$, we conclude $\widehat{\mathbf{x}} = \Pi^p_{\mathcal{M}_{\mathbf{Ax=y}}}(\mathbf{x}_0)$. As we have shown that for every subsequence of $\{\mathbf{x}_n\}_n$ there is a subsequence that in turn converges weakly to the same limit $\Pi^p_{\mathcal{M}_{\mathbf{Ax=y}}}(\mathbf{x}_0)$, then this holds for the sequence $\{\mathbf{x}_n\}_n$, too, see [29, Sect. 10.5], and we are done. □

*Remark* 1. Note that the duality mappings of $\ell_p$-spaces, where $1 < p < \infty$, are sequentially weak-to-weak-continuous, see [4, Remark 4.3], and hence satisfy the assumptions of Theorem 3.

If we choose $\mathbf{x}_0 = 0$, then $\widehat{\mathbf{x}} = \mathbf{x}^\dagger$ is the minimum-norm-solution of (2).

### *3.5. Orthogonalized SESOP as generalized CGNE*

Of special importance and interest is the connection between the orthogonalized SESOP and the CG method applied to the normal equation in Hilbert spaces.

The method of conjugate gradients was first proposed by [5] and nowadays a whole zoo of CG methods exist, see [30] for a taxonomoy. It is especially competitive for problems with a large discretization dimension where a direct inversion or second-order methods such as BFGS are prohibitively expensive. Standard CG requires a symmetric, positive definite operator $\mathbf{A}$ in a Hilbert space $\mathcal{X}$. However, it is also applicable in the case of a non-symmetric operator, i. e. $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{X} \neq \mathcal{Y}$, if it is applied to the normal equation $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{y}$, following [31, Sect. 8.1, 8.3], which is equivalent to minimizing the residual $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}$. This leads to the so called *Conjugate Gradient Normal Residual* (CGNR) method, which is used in the context of over-determined systems with injective $\mathbf{A}$.

Alternatively, for underdetermined systems we can investigate

$$\mathbf{A}\mathbf{A}^*\mathbf{b} = \mathbf{y} \quad \text{with } \mathbf{x}^\dagger = \mathbf{A}^* b$$

and thus compute the solution with minimal norm

$$\left\|\mathbf{x}^\dagger\right\| = \min\left\{\|\mathbf{x}\| : \mathbf{A}\mathbf{x} = \mathbf{y}\right\} \tag{46}$$

which is called *Conjugate Gradient Normal Equation* (CGNE) or *Craig's method*, see [11].

Let us first state some identities if $\mathcal{X}$, $\mathcal{Y}$ are Hilbert spaces and $p = 2$,

- $\mathcal{X} = \mathcal{X}^*$, $\mathcal{Y} = \mathcal{Y}^*$,
- $\Delta_2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$,
- $\Pi^2_{\mathcal{C}}(\mathbf{x}) = P_{\mathcal{C}}(\mathbf{x})$, $\mathcal{C} \subset \mathcal{X}$ being closed and convex,
- $J^{\mathcal{X}}_2 = \mathbb{1}_{\mathcal{X}}$, $j^{\mathcal{Y}}_2 = J^{\mathcal{Y}}_2 = \mathbb{1}_{\mathcal{Y}}$.

Hence, Method 1 calculates the metric projection onto the solution manifold, see Theorem 3. Hence, it minimizes the norm $\|\mathbf{z}\|$ over all $\mathbf{z} \in \mathcal{M}_{\mathbf{Ax=y}}$. According to [32, Algorithm 2.2] or [30, Sect. 2.3] with $\mathbf{B} = \mathbb{1}$, the update in the CGNE method is

$$\begin{aligned}
r_0 &= \mathbf{y} - \mathbf{A}\mathbf{x_0}, \quad \mathbf{p_0} = \mathbf{A}^*\mathbf{r_0} \\
\alpha_n &= \frac{\langle \mathbf{r_n}, \mathbf{r_n} \rangle}{\langle \mathbf{p_n}, \mathbf{p_n} \rangle} \\
\mathbf{x_{n+1}} &= \mathbf{x_n} + \alpha_n \mathbf{p_n}, \quad \mathbf{r_{n+1}} = \mathbf{r_n} - \alpha_n \mathbf{A}\mathbf{p_n} \\
\mathbf{p_{n+1}} &= \mathbf{A}^*\mathbf{r_{n+1}} + \beta_n \mathbf{p_n}, \quad \beta_n = -\frac{\langle \mathbf{A}^*\mathbf{r_{n+1}}, \mathbf{p_n} \rangle}{\langle \mathbf{p_n}, \mathbf{p_n} \rangle}
\end{aligned}$$

We immediately see the equivalence with the update scheme of Method 1 using a single search direction when substituting as follows,

$$\begin{aligned}
\mathbf{r_n} &\to -\mathbf{R}_n = \mathbf{A}\mathbf{x}_n - \mathbf{y}, \\
\mathbf{p_n} &\to \mathbf{v}_n^* = \mathbf{A}^*\mathbf{R}_n - s_n \mathbf{v}_{n-1}^*, \\
\alpha_n &\to \mu_n = \frac{\langle \mathbf{o}_n^*, \mathbf{R}_n \rangle}{\langle \mathbf{v}_n^*, \mathbf{v}_n^* \rangle} = \frac{\langle \mathbf{R}_n - s_{n-1}\mathbf{o}_{n-1}^*, \mathbf{R}_n \rangle}{\langle \mathbf{v}_n^*, \mathbf{v}_n^* \rangle} = \frac{\langle \mathbf{R}_n, \mathbf{R}_n \rangle}{\langle \mathbf{v}_n^*, \mathbf{v}_n^* \rangle} \\
\beta_{n-1} &\to s_n = \frac{\langle \mathbf{v}_{n-1}^*, \mathbf{d}_n^* \rangle}{\langle \mathbf{v}_{n-1}^*, \mathbf{v}_{n-1}^* \rangle} = \frac{\langle \mathbf{v}_{n-1}^*, \mathbf{A}^*\mathbf{R}_n \rangle}{\langle \mathbf{v}_{n-1}^*, \mathbf{v}_{n-1}^* \rangle},
\end{aligned}$$

where we used (20) and (30). In that sense SESOP with orthogonalized search directions can be seen as a generalization of the CGNE method to Banach spaces. Using more than one previous search direction corresponds to reorthogonalization when conjugacy is lost due to rounding errors; see also [33, 34, 35], where furthermore the relationship between the conjugate gradient method and the Lanczos method is discussed.

## 4. Numerical experiments

We perform numerical experiments that compare Method 1 and Method 2, namely SESOP using either the unorthogonalized search space $\mathcal{U}_n^{\text{trunc}}$ or the orthogonalized search space $\mathcal{V}_n^{\text{trunc}}$. In the following we refer to both methods simply by the corresponding search space that each one employs. The first part uses the problem from [4, Sect. 5], namely solving $\mathbf{A}\mathbf{x} = \mathbf{y}$ for various $\ell_p$ spaces and uniformly distributed random matrices $\mathbf{A}$ and right-hand sides $\mathbf{y}$. In the second part we solve inverse problems in 2D computerized tomography.

All experiments have been performed on a single core of an Intel Xeon E3-1270 cpu with 3.50GHz. Note that SESOP and the orthogonalization procedure described in Section 3 have been implemented in the C++ library *BASSO* (*BA*nach *S*equential *S*ubspace *O*ptimizer), based on the Eigen3 library [36] for the linear algebra routines[2].

### *4.1. Toy problem*

We first look at the inverse problem of a random matrix and a random right-hand side to be formally inverted as is done frequently in the literature, but with well-known shortcomings, see [37]. In this experiment, we want to check the case of $p = 2$ for a single and for multiple search directions. We will see that with SESOP there is still a significant decrease in required iterations from single to multiple search direction for the $\ell_2$ space. With CG no such difference arises due to the conjugacy property. Furthermore, to assess a possible speed-up of the orthogonalized search directions, we investigate various $\ell_p$ spaces and norms.

#### *4.1.1. Procedure*

To this end, we create a uniformly distributed random matrix $\mathbf{A} \in [-1, 1]^{l\times m}$ with $l = 1000$ and $m = 5000$, representing a discretized version of some random operator. Next, we want to create a solution $\mathbf{x}^\dagger \in \mathbb{R}^m$ to a random right-hand side $\mathbf{y} \in \mathbb{R}^l$ in the sense that this solution should be a minimum-norm solution in an $\ell_p$-space with $p \in \{1.1, 1.2, 1.5, 2, 3, 6, 10\}$. Therefore, we create a random right-hand side precursor $\mathbf{y}^* \in [-1, 1]^l$ and calculate the minimum-norm solution $\mathbf{x}^\dagger$ as follows:

$$\mathbf{x}^\dagger := \frac{J^*_{p^*}(\mathbf{A}^* y^*)}{\left\| J^*_{p^*}(\mathbf{A}^* y^*) \right\|_p} \tag{47}$$

[2] BASSO is available from GitHub at https://github.com/FrederikHeber/BASSO under GPLv2 license.

Then we finally obtain the right-hand side as $\mathbb{R}^l \ni \mathbf{y} = \mathbf{A}\mathbf{x}^\dagger$. We use 10 different seeds $s \in \{420, \ldots, 429\}$ for the random number generator and calculate average iteration counts $n$ and standard deviations $\sigma_n$ over these 10 runs with otherwise identical parameters. The iteration is stopped at either $n > 20,000$ or if the relative residual $\frac{\|\mathbf{A}\mathbf{x}_n - \mathbf{y}\|_\mathcal{Y}}{\|\mathbf{y}\|_\mathcal{Y}}$ is less than $10^{-4}$. Note that we always set $\mathcal{Y} = \ell_2(\mathbb{R}^l)$.

### *4.1.2. SESOP*

First, we reproduce the results from [4, Sect. 5] to elucidate any possible differences that arise from different implementations, see Table 1. There, the matrix dimension is $1000 \times 5000$, the $\ell_p$-norms of $\mathcal{X}$ are given for $p \in \{1.2, 1.5, 6, 10\}$, the power of the gauge function of the duality mapping $J_p$ is given as $p$ for $p \geq 2$ and $p = 2$, else. We use $N \in \{2, 4, 6\}$ numbers of search directions.

If we take into account that in Table 1(a) only a single run is given and comparing this to our averages and standard deviations, the iteration counts are in very good agreement up until $p = 2$. For larger $p$ the discrepancy is quite large. However, this is also true for the standard deviations. Hence, overall we do not see any discrepancy resulting from the different implementations, i. e. we have a solid base for comparing the results with the original MatLab implementation of [4].

### *4.1.3. Orthogonalized SESOP*

Next, we have a look at the change in iteration counts and runtimes between the search space $\mathcal{U}_n^{\text{trunc}}$ used in SESOP and the search spaces $\mathcal{V}_n^{\text{trunc}}$ using metric projections as proposed in this article. We use various numbers of search directions $N \in \{1, 2, 4, 6\}$ and $\ell_p$-spaces with $p \in \{1.1, 1.2, 1.5, 2, 3, 6, 10\}$ and the power type of the duality mapping chosen as in Subsection 4.1.2. For comparison we also implemented a Landweber method where the step size is chosen such that the residual is minimized along the current Landweber search direction. For this optimization problem we employ *Brent's method*, i. e. we do not use any gradient information. The average iteration counts and standard

**Table 1:** Comparison of SESOP implementations from [4] and this work (without orthogonalization): Average iteration counts $n$ and standard deviations $\sigma_n$ for various $p$ and $N$ values and matrix dimension $m = 5000$.

(a) from [4]

| p | N | $n$ |
|---|---|---|
| 1.2 | 2 | 435 |
| 1.2 | 4 | 211 |
| 1.2 | 6 | 137 |
| 1.5 | 2 | 22 |
| 1.5 | 4 | 15 |
| 1.5 | 6 | 14 |
| 6 | 2 | 102 |
| 6 | 4 | 79 |
| 6 | 6 | 49 |
| 10 | 2 | 297 |
| 10 | 4 | 183 |
| 10 | 6 | 131 |

(b) This work

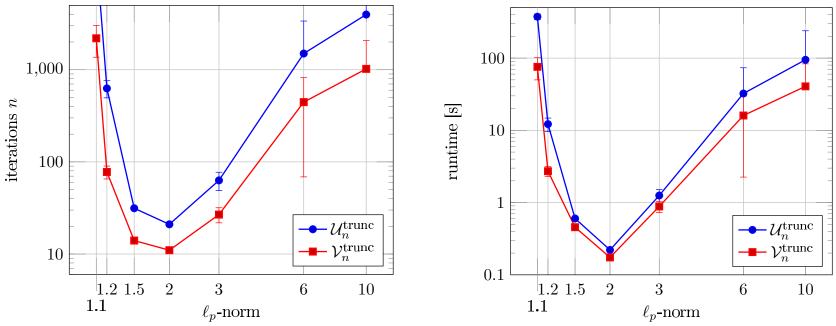
| p | N | $n$ | $\sigma_n$ |
|---|---|---|---|
| 1.2 | 2 | 402.9 | 80.75 |
| 1.2 | 4 | 199.1 | 35.72 |
| 1.2 | 6 | 133.2 | 19.59 |
| 1.5 | 2 | 21.7 | 0.46 |
| 1.5 | 4 | 14.9 | 0.3 |
| 1.5 | 6 | 14 | 0 |
| 6 | 2 | 691.7 | 744.38 |
| 6 | 4 | 200.1 | 90.35 |
| 6 | 6 | 99.2 | 31.99 |
| 10 | 2 | 1,413.8 | 1,399.17 |
| 10 | 4 | 396.6 | 195.55 |
| 10 | 6 | 188.3 | 64.18 |

deviations are given in Table 2.

We notice that with $p \in \{1.2, 1.5, 2\}$ the average iteration count $n$ changes only slightly with respect to different number of search directions $N$. This holds for $p = 1.5$ within output precision of $10^{-7}$ and for $p = 2$ within full floating point numerical precision. Hence, we see that if $\mathcal{X} = \ell_2(\mathbb{R}^m)$ is a Hilbert-space, then the orthogonality is maintained between all search directions from $\mathbf{v}^*_{0,N_0}$ up to $\mathbf{v}^*_{n,N_n}$. This is the behavior expected from CG methods and we elucidate this further in the next section. It is maintained to some extent also if $p$ is close to 2.

Additionally, for $p \in \{1.5, 2\}$ we observe that for a large number of search directions $N$ the average number of iteration steps $n$ is similar for both of the search spaces $\mathcal{U}_n^{\text{trunc}}$ and $\mathcal{V}_n^{\text{trunc}}$. This indicates that using more than one search direction, the central idea of [4], the subspaces $\mathcal{U}_n^{\text{trunc}}$ and $\mathcal{V}_n^{\text{trunc}}$ at each step $n$ tend to be identical if the number of search directions goes to infinity, namely $\mathcal{U}^{\text{exp}} = \mathcal{V}^{\text{exp}}$, c.f. Section 3.5.

At last, iteration counts and deviations become very large for $p \to 1$ and $p \to \infty$, i. e. when the $\ell_p$ spaces are no longer smooth.



(a) Iterations

(b) Runtimes

**Figure 2:** Iterations and runtimes using SESOP and two different search spaces for the minimum-norm solution of $\mathbf{A}\mathbf{x} = \mathbf{y}$ with uniformly random $\mathbf{A} \in \mathbb{R}^{1000\times5000}$ and a single search direction.

In Figure 2 we depict for a single search direction both iteration counts and

**Table 2:** Averaged iteration counts $n$ and standard deviations $\sigma_n$ for various $p$ and $N$ values and $m = 5000$ for solving with SESOP for the minimum-norm solution of $\mathbf{Ax} = \mathbf{y}$: *unorth*ogonalized refers to $\mathcal{U}^{\text{trunc}}$ and *metric* to $\mathcal{V}^{\text{trunc}}$.

| p | N | $n^{\text{unorth}}$ | $\sigma_n^{\text{unorth}}$ | $n^{\text{metric}}$ | $\sigma_n^{\text{metric}}$ |
|---|---|---|---|---|---|
| 1.1 | 1 | 18,880.6 | 1,533.87 | 2,196.4 | 828.39 |
| 1.1 | 2 | 13,739.7 | 3,078.62 | 1,243 | 252.83 |
| 1.1 | 4 | 7,167.5 | 1,004.05 | 815.6 | 67 |
| 1.1 | 6 | 4,187.3 | 509.16 | 654.4 | 65.69 |
| 1.2 | 1 | 627.8 | 132.57 | 77.7 | 12.4 |
| 1.2 | 2 | 402.9 | 80.75 | 77.2 | 8.33 |
| 1.2 | 4 | 199.1 | 35.72 | 71 | 7.56 |
| 1.2 | 6 | 133.2 | 19.59 | 67.8 | 6.43 |
| 1.5 | 1 | 31.4 | 0.8 | 14 | 0 |
| 1.5 | 2 | 21.7 | 0.46 | 14 | 0 |
| 1.5 | 4 | 14.9 | 0.3 | 14 | 0 |
| 1.5 | 6 | 14 | 0 | 14 | 0 |
| 2 | 1 | 21 | 0.45 | 11 | 0 |
| 2 | 2 | 14.9 | 0.3 | 11 | 0 |
| 2 | 4 | 11.3 | 0.46 | 11 | 0 |
| 2 | 6 | 11 | 0 | 11 | 0 |
| 3 | 1 | 63 | 14.2 | 26.8 | 5 |
| 3 | 2 | 43.5 | 8.8 | 21.9 | 3.81 |
| 3 | 4 | 22.1 | 2.74 | 18.7 | 1.68 |
| 3 | 6 | 19.3 | 1.68 | 18.4 | 1.36 |
| 6 | 1 | 1,499.3 | 1,879.2 | 445.3 | 376.56 |
| 6 | 2 | 691.7 | 744.38 | 262.4 | 183.25 |
| 6 | 4 | 200.1 | 90.35 | 99.2 | 38.2 |
| 6 | 6 | 99.2 | 31.99 | 56.9 | 14.94 |
| 10 | 1 | 3,973.9 | 5,692.81 | 1,021.7 | 1,049.87 |
| 10 | 2 | 1,413.8 | 1,399.17 | 563.4 | 507.32 |
| 10 | 4 | 396.6 | 195.55 | 208.1 | 82.09 |
| 10 | 6 | 188.3 | 64.18 | 112 | 33.12 |

the total runtime for solving for the minimum-norm solution up to a relative residual threshold of $10^{-4}$ or up to $20,000$ iteration steps. Here, we want to compare the method's performance with either search space directly. We notice that iteration counts for the orthogonalized search space $\mathcal{V}_n^{\text{trunc}}$ are at least a factor of 2-3 below the ones for the search space $\mathcal{U}_n^{\text{trunc}}$. This holds for all values of $p$. This reduced number of iterations, that are necessary for the same residual threshold, is the reason, why the runtimes show the same trend between the different search spaces up to a similar factor, despite the additional computational effort for the orthogonalization.

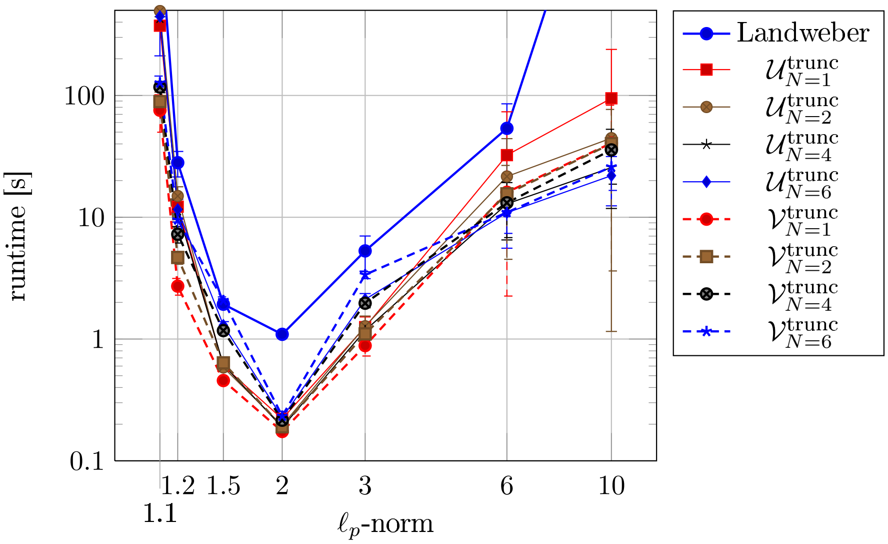


**Figure 3:** Runtimes using Landweber and SESOP with unorthogonalized and orthogonalized search spaces for the minimum-norm solution of $\mathbf{A}\mathbf{x} = \mathbf{y}$ with uniformly random $\mathbf{A} \in \mathbb{R}^{1000\times 5000}$ and $N \in \{1, 2, 4, 6\}$.

Finally, we look at runtimes for an increasing number of search directions in Figure 3. There, we compare Landweber using the dynamic step size with SESOP and either search space $\mathcal{U}_n^{\text{trunc}}$ or search space $\mathcal{V}_n^{\text{trunc}}$. We notice that the orthogonalized search space $\mathcal{V}_{N=1}^{\text{trunc}}$ using a single search direction is fastest for $p \leq 3$. For larger $p \geq 6$ we see that the search spaces $\mathcal{U}_{N=6}^{\text{trunc}}$ and $\mathcal{V}_{N=6}^{\text{trunc}}$, both using the highest number of investigated search directions, are comparable with respect to runtime. However, we also see that using more search directions $N$ significantly slows down SESOP using $\mathcal{V}_N^{\mathrm{trunc}}$ for $p \leq 3$. This is much less pronounced with search space $\mathcal{U}_N^{\mathrm{trunc}}$ where no additional cost for the orthogonalization is accumulated.

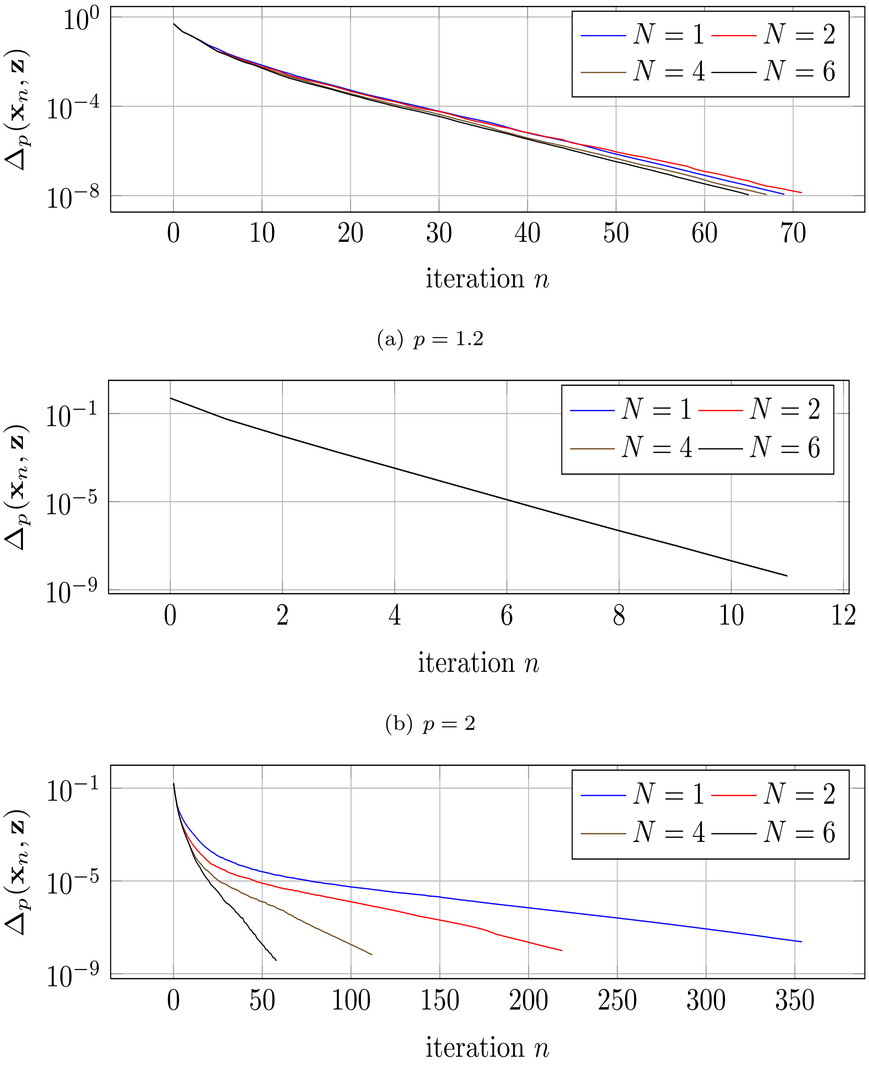
The results can be explained as follows: On the one hand, for small $p$ the additional cost of the orthogonlization procedure for many search directions outweighs any advantage obtained by converging in fewer iterations. There are two reasons behind this. First, SESOP requires less iteration steps when using multiple search directions, c. f. Table 2. Second, the overhead of orthogonalization becomes more costly when $N$ increases. On the other hand, for very large $p$ and, because of the nature of the associated dual space, also for $p$ close to 1, the orthogonalization is numerically very difficult due to the large powers involved in the $\ell_p$-norms. Finally, Landweber's method with a residual minimizing step size is always an order of magnitude slower than the fastest subspace method. We note that the Landweber method with fixed stepwidth $\mu_n$ according to [1, Method 3.1] is already for the $\ell_2$ norm very far from being competitive. It requires on average $42300 \pm 750$ iteration steps, the runtime around $207 \pm 12$ seconds. This is slower than the fastest subspace method for any of the investigated $\ell_p$-norms.

As a general rule we summarize that orthogonalized SESOP with search space $\mathcal{V}_n^{\mathrm{trunc}}$ is fastest with a single search direction for small $p \leq 3$ and with multiple search directions for $p \geq 6$.

### *4.1.4. Connection to the CG method in Hilbert space*

We already discussed the connection of Method 1 to the CGNE method in Section 3.5. According to [30, Table 5.2], the inner product matrix of CGNE is the identity matrix. Hence, the conjugacy is reduced to a simple orthogonality of search directions. Numerically, we therefore expect that multiple search directions $N$ do not change the convergence behavior in any way for $\ell_2$-spaces.

In Figure 4 we visualize the error with respect to the Bregman distance for 3 different $\ell_p$ norms. There, we only look at the orthogonalized search space



(a) $p = 1.2$

(b) $p = 2$

(c) $p = 6$

**Figure 4:** Bregman distance $\Delta_p(\mathbf{x}_n, \mathbf{z}) = \|\mathbf{x}_n - \mathbf{z}\|^2$ for $\mathcal{X} = \ell_2(\mathbb{R}^m)$, $p = 2$, depicted over the iteration number $n$ for SESOP with search space $\mathcal{V}_n^{\text{trunc}}$ for increasing number of search directions $N$, for various $\ell_p$ norms with $m = 5000$, and for a specific random number generator seed.

$\mathcal{V}_n^{\text{trunc}}$ using metric projections. We clearly recognize that in the $\ell_2$-case more than a single search direction does not change anything about the minimization. It proceeds (up to numerical precision) in exactly the same manner as if there were only a single search direction spanning the search space. This is precisely what would be expected of a CG method due to the inherent conjugacy property. This is not valid for the other two cases $p = 1.2$ and $p = 6$ where more search directions lead to a significant speed-up, especially for $p \gg 2$. Furthermore, we see in Table 2 that there is no difference between a single search direction and multiple search directions as the current gradient direction is made orthogonal to all previous search directions simultaneously. As a consequence we may conclude by this numerical example that the subspace methods (with orthogonalization) to a certain degree in fact are an extension of the CG methods to general Banach spaces.

### *4.2. Computerized Tomography*

The problem of computerized tomography aims to reconstruct the interior of an object from projections. We refer to the seminal books [38, 39] for the mathematics of computerized tomography. Measurements are obtained by passing X-rays through a body, whose intensity is diminished, proportional to the passed length and density of the body $f(x) : [0,1]^2 \to \mathbb{R}_0^+$. This decrease is measured over $a$ angles and $s$ shifts of radiation source and detectors. The arising measurement matrix is usually called *sinogram*. We follow [40, Sect. 7.7] for a brief introduction to the discretization of the problem.

The measurement rays are parametrized as

$$t^i(\tau) = t^{i,0} + \tau d^i, \tag{48}$$

where $d$ is the directional vector of the $i$th ray with $i = 1, \ldots, s \cdot a$. Then, the attenuation of the $i$-th ray can be written as the *Radon transform*,

$$b_i = \int_{-\infty}^{\infty} f\left(t^i(\tau)\right) d\tau, \quad i = 1, \ldots, s \cdot a, \tag{49}$$

i. e. we integrate $d\tau$ along the line $t^i(\tau)$. The problem can be discretized using a pixel basis,

$$\chi_{kl}(x) = \begin{cases} 1, & x \in [h \cdot (k-1), h \cdot k] \times [h \cdot (l-1), h \cdot l] \\ 0, & \text{else} \end{cases}, \quad \text{for all } k, l = 1, \ldots, M, \tag{50}$$

where we assume the absorption coefficient $f(x) = \sum_{kl} f_{kl} \chi_{kl}(x)$ to be piecewise constant with $h = \frac{1}{M}$ and $f_{kl}$ is the grey-scale value at pixel $\chi_{kl}$ to be computed.

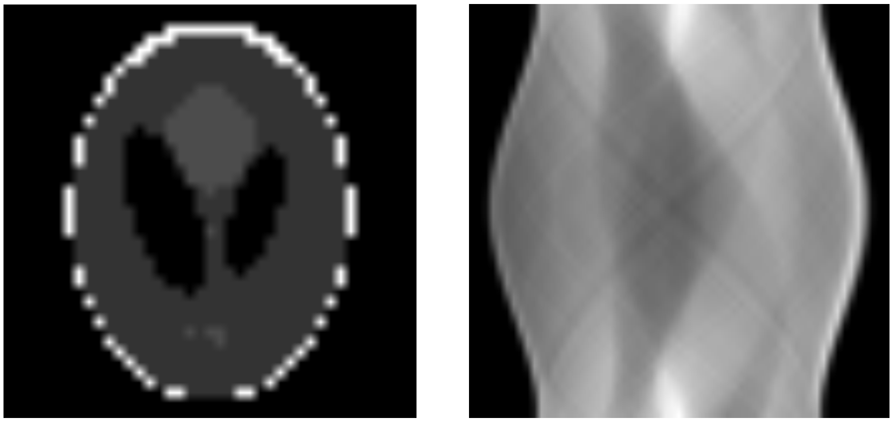
For the discretization of (49) we simply need to count the length $\Delta L_{kl}^{(i)}$ of each ray $t^i$ in each pixel $\chi_{kl}$ of the basis (50) and obtain

$$b_i = \sum_{k,l=1}^{M} f_{kl} \Delta L_{kl}^{(i)} \quad \text{for } i = 1, \ldots, s \cdot a. \tag{51}$$

If we vectorize the matrix object $f_{kl}$ to become the vector $x_j$ with $j = (l-1)M + k$, then we obtain

$$b_i = \sum_j a_{ij} x_j, \quad i = 1, \ldots, s \cdot a. \tag{52}$$

Note that the matrix $A$ is sparse which we exploit in the implementation.



(a) Phantom

(b) Sinogram

**Figure 5:** Shepp-Logan phantom and its sinogram.

As benchmark we take the standard Shepp-Logan phantom, see Figure 5(a), where we obtain the measurements by using the known analytical Radon transform of ellipses, see Figure 5(b). We discretize using $M = 41$ pixels, $s = 61$ number of shifts, and $a = 60$ number of angles. We deliberately choose a coarser image resolution together with a higher number of measurements to allow for a high-quality reconstruction and clearly discernable artifacts if there are any. Note that we additionally project the solution onto the range of the matrix $A$ such that an exact solution exists. We return to this point in Section 4.2.2.

We again want to compare SESOP using the truncated search space $\mathcal{U}_n^{\text{trunc}}$ as well as orthogonalized SESOP with search space $\mathcal{V}_n^{\text{trunc}}$. The dimensions are as follows: $|\mathcal{U}_n^{\text{trunc}}| = 2$ and $|\mathcal{V}_n^{\text{trunc}}| = 1$. We stop the methods either after 500 iterations or if the absolute residual is less than $10^{-2}$. We use $\ell_p$-spaces with $p \in \{1.1, 1.2, 1.5, 2\}$ for $\mathcal{X}$ and $\ell_2$ for $\mathcal{Y}$. For the line-search problem we use at most 20 iterations. The power type of the duality mappings is always set to $p = 2$.

### *4.2.1. Exact data*

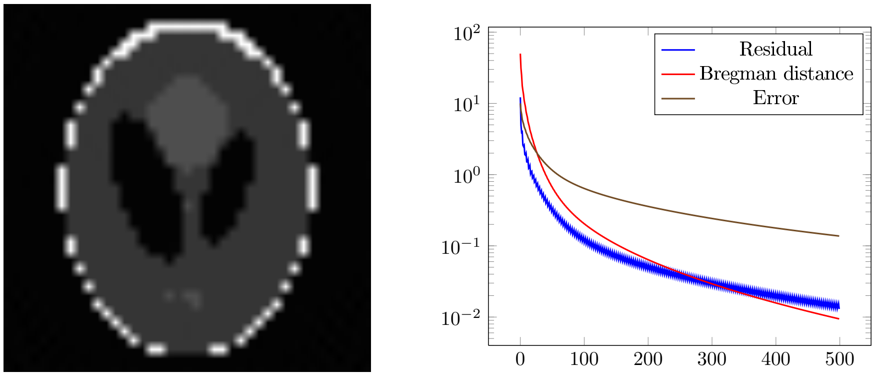
First of all, we study the situation of exact data and $\mathcal{X} = \ell_2$, i. e. a Hilbert space setting. Figure 6 displays reconstructions along with the error, residual, and Bregman distance histories for SESOP with the truncated search space $\mathcal{U}_n^{\text{trunc}}$ and two search directions.

We realize that SESOP converges but stops at $n = 500$ iterations where the absolute residual is still slightly larger than $10^{-2}$. The overall runtime is 2.4 seconds. Note that the convergence is monotone only in the Bregman distance and not in the residual, c.f. Theorem 3.

Next, we apply orthogonalized SESOP using the search space $\mathcal{V}_n^{\text{trunc}}$ with a single search direction[3]. We obtain the results depicted in Figure 7.

We notice that the iteration stops at roughly $n = 60$ when reaching an

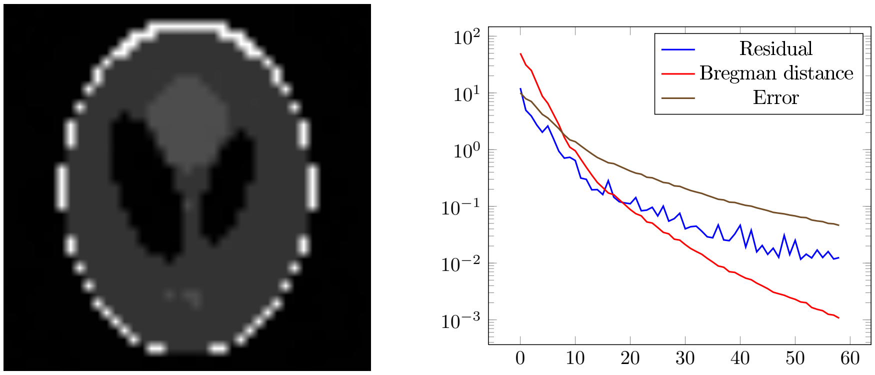
[3] As mentioned before, in $\ell_2$ more search directions do not improve performance when using an orthogonalized search space.



(a) solution

(b) iteration history

**Figure 6:** Recovered solution and residual, error, and Bregman distance histories using SESOP with $\left|\mathcal{U}_n^{\text{trunc}}\right| = 2$ directions.



(a) solution

(b) iteration history

**Figure 7:** Recovered solution and residual, error, and Bregman distance histories using orthogonalized SESOP with $\left|\mathcal{V}_n^{\text{trunc}}\right| = 1$ direction.

absolute residual of $10^{-2}$. Naturally, this leads to a faster runtime compared to the unorthogonalized search space of only 0.36 seconds. In the former case the reconstructed image shows some very slight artifacts, in the latter case no artifacts are visible.

### *4.2.2. Noisy data*

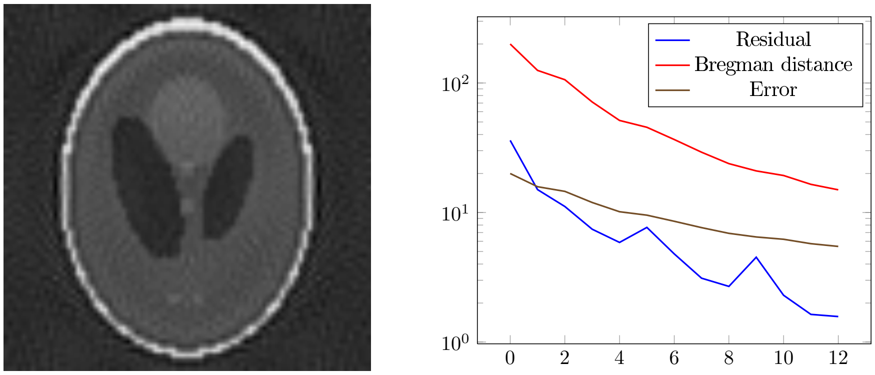
It is called *inverse crime*, if the same discretization is being used for both, the operator and the right-hand side, see [40, Sect. 7.2]. Results may look suspiciously good in this case, i. e. no mismatch between (real) data and model is revealed. We definitely committed this crime by projecting our measurements onto the range of the matrix $\mathbf{A}$. This is why in the next step we additionally disturb the right-hand side $\mathbf{y}$, obtained from projecting the Shepp-Logan phantom $\mathbf{x}^\dagger$, where $\mathbf{y} = \mathbf{A}\mathbf{x}^\dagger$, with noise of known level $\delta$ to get a contaminated data vector $\mathbf{y}^\delta$.

To this end, we compute a random vector $\mathbf{n} \in [-1, 1]^m$ and set $\widetilde{\mathbf{y}} = \mathbf{y} + \delta \frac{\|\mathbf{y}\|}{\|\mathbf{n}\|} \mathbf{n}$. We use a noise level of $\delta = 0.01$. We stop the iteration, if the relative residual threshold is less than 0.03, i.e. we use a tolerance parameter of factor 3. We increase the number of offsets to $s = 81$ and number of angles to $a = 80$. We furthermore increase the number of pixels $M = 81$ of the reconstructed image. We still remain in the Hilbert space setting $\mathcal{X} = \ell_2$ for the moment. In the next section we consider other $\ell_p$ spaces.

In Figure 8 both, the reconstructed image and iteration history with residual, Bregman distance, and error with respect to the true solution is shown. We obtain good results with respect to the noise level employed. The runtime is 0.34 seconds. We conclude that we do not commit any inverse crime and that the proposed method is indeed working and implemented in a decent way.

### *4.2.3. Small $\ell_p$ Norms*

Finally, we also look at other norms than $\ell_2$, namely small $\ell_p$ norms with $p \in \{1.1, 1.2, 1.5\}$. We again use $s = 61$, $a = 60$ and $M = 41$ to allow for a high-quality reconstruction.



(a) solution

(b) iteration history

**Figure 8:** Recovered solution (a), residual, error, and Bregman distance histories (b) using orthogonalized SESOP with $\left|\mathcal{V}_n^{\text{trunc}}\right| = 1$ direction in the presence of noise of level 0.01 and a tolerance parameter 3.

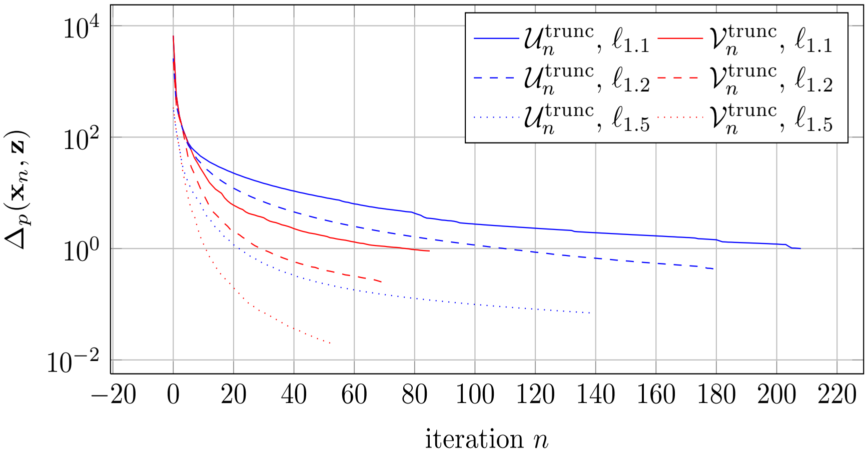
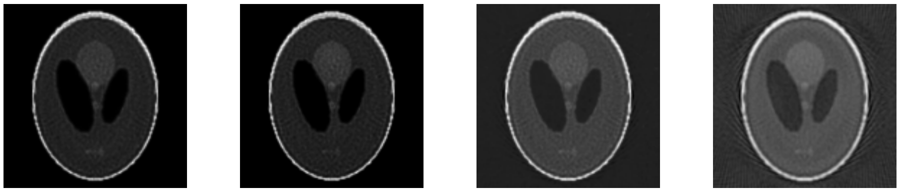


**Figure 9:** Bregman distance $\Delta_p(\mathbf{x}_n, \mathbf{z})$ depicted over the iteration number $n$ using SESOP with unorthogonalized search space $\mathcal{U}_n^{\text{trunc}}$ and orthogonalized search space $\mathcal{V}_n^{\text{trunc}}$ for spaces norms $\ell_{1.1}$, $\ell_{1.2}$, and $\ell_{1.5}$.

In Figure 9 we compare the decrease in Bregman distance for unorthogonalized and orthogonalized search spaces for various small $\ell_p$ norms. We see that for all tested small $\ell_p$ norms the orthogonalized search space $\mathcal{V}_n^{\text{trunc}}$ requires about three times fewer iterations to reach the stopping criterion. This makes it also faster in the overall runtimes, e. g. for $p = 1.1$ we obtain 3.15 seconds without orthogonalization and 2.6 seconds when employing $\mathcal{V}_n^{\text{trunc}}$. Note that the method terminates at different Bregman distances, since different $\ell_p$-norms have been used to measure the residual stopping criterion.



(a) $\ell_{1.1}$ (b) $\ell_{1.2}$ (c) $\ell_{1.5}$ (d) $\ell_2$

**Figure 10:** Reconstructed Shepp-Logan phantom in the presence of noise with level $\delta = 0.01$ and using small $\ell_p$ norms with $p \in \{1.1, 1.2, 1.5, 2\}$.

We show in Figure 10 larger reconstructed images for $M = 127$ pixels with a noise of $\delta = 0.01$ while using the same number of angles $a$ and shifts $s$ than before. This is to elucidate the effect of the different $\ell_p$ norms in the presence of noise. We observe that for smaller $\ell_p$ both contrast and noise of the image is enhanced. Especially, artifacts surrounding the reconstructed phantom in the $\ell_2$ case are absent for $\ell_{1.1}$. On the other hand, noisy speckles are more present in the reconstructed image using the $\ell_{1.1}$ norm.

## 5. Conclusions

Based on the previous work of [10] and on the concept of orthogonality by [12], we have proposed an orthogonalized set of search directions in Banach spaces using metric projections. Using search spaces consisting only of a finite number of Landweber directions modified by this orthogonalization, we have shown that the SESOP method converges weakly and that in Hilbert spaces the procedure coincides with the CGNE method, also known as Craig's method [11]. In this respect, the subspace methods can be seen as a natural extension of CG methods to general Banach spaces.

Numerical experiments have shown fast convergence for both an inverse toy problem consisting of a uniformly distributed random matrix and right-hand side on various $\ell_p$-spaces as well as for the inverse problem of 2D computerized tomography for reconstructing the Shepp-Logan phantom. In every case the orthogonalized truncated search space clearly outperforms the truncated search space used in [4] with respect to both required numbers of iterations and runtime. Note that either method is faster by at least an order of magnitude than Landweber's method with a gradient-free line search.

As an outlook, we would like to remind that there is a whole zoo of CG-variants, see [41, p. 98]. It would be very insightful to find more connections between a specific variant and a choice of (orthogonalized) search spaces for the orthogonalized SESOP. The regularizing property of CG methods in Hilbert spaces is well-known due to [42, 43], its generalization to Banach spaces is an open question. Finally, since there are also CG variants for non-linear problems and SESOP has been extended to nonlinear, ill-posed problems in Hilbert spaces [9], this would be a very interesting research field for the subspace methods as well.

**Acknowledgements**

The authors acknowledge funding by the German Ministry of Education and Research (BMBF) under the project HYPERMATH (05M2013).

## References

[1] F. Schöpfer, A. K. Louis, T. Schuster, Nonlinear iterative methods for linear ill-posed problems in Banach spaces, Inverse Problems 22 (1) (2006) 311–329. doi:10.1088/0266-5611/22/1/017.

[2] M. Hanke, A. Neubauer, O. Scherzer, A convergence analysis of the Landweber iteration for nonlinear ill-posed problems, Numerische Mathematik 72 (1) (1995) 21–37.

[3] A. Kirsch, An Introduction to the Mathematical Theory of Inverse Problems, 2nd Edition, Springer Science+Business Media, New York Dordrecht Heidelberg London Library, 2011. doi:10.1007/978-1-4419-8474-6.

[4] F. Schöpfer, T. Schuster, A. K. Louis, Metric and bregman projections onto affine subspaces and their computation via sequential subspace optimization methods, Journal of Inverse and Ill-posed Problems 16 (5) (2008) 1–29. doi:10.1515/JIIP.2008.026.

[5] M. R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards 49 (6) (1952) 409–436. doi:10.6028/jres.049.044.

[6] J. Nocedal, S. J. Wright, Numerical Optimization, 1st Edition, Springer-Verlag, New York, 1999.

[7] G. Narkiss, M. Zibulevsky, Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems (2005).

[8] F. Schöpfer, T. Schuster, Fast regularizing sequential subspace optimization in Banach spaces, Inverse Problems 25 (1) (2009) 1–22. doi:10.1088/0266-5611/25/1/015013.

[9] A. Wald, T. Schuster, Sequential subspace optimization for nonlinear inverse problems, Journal of Inverse and Ill-Posed Problems 25 (4). doi:10.1515/jiip-2016-0014.

[10] F. Schöpfer, Iterative Regularization Methods for the Solution of the Split Feasibility Problem in Banach Spaces, Doctoral thesis, Universität des Saarlandes (2007).

[11] E. J. Craig, The N -Step Iteration Procedures, Journal of Mathematics and Physics 34 (1-4) (1955) 64–73. doi:10.1002/sapm195534164.

[12] Y. I. Alber, James orthogonality and orthogonal decompositions of banach spaces, Journal of Mathematical Analysis and Applications 312 (1) (2005) 330–342. doi:10.1016/j.jmaa.2005.03.027.

[13] I. Cioranescu, Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems, 1st Edition, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[14] T. Schuster, B. Kaltenbacher, B. Hofmann, K. S. Kazimierski, Regularization Methods in Banach Spaces, De Gruyter, 2012.

[15] J. Lindenstrauss, On the modulus of smoothness and divergent series in Banach spaces., The Michigan Mathematical Journal.

[16] J. A. Clarkson, Uniformly Convex Spaces, Transactions of the American Mathematical Society 40 (3) (1936) 396. doi:10.2307/1989630.

[17] M. M. Day, Uniform Convexity in Factor and Conjugate Spaces, The Annals of Mathematics 45 (2) (1944) 375. doi:10.2307/1969275.

[18] J. Lindenstrauss, L. Tzafriri, Classical Banach Spaces II - Function Spaces, 1st Edition, Springer-Verlag Berlin Heidelberg New York, 1979.

[19] M. Schechter, Principles of functional analysis, 1st Edition, Academic Press, Inc., New York, 1971.

[20] Z.-b. Xu, G. F. Roach, Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces, Journal of Mathematical Analysis and Applications 157 (1) (1991) 189–210. doi:10.1016/0022-247X(91)90144-O.

[21] C. Chidume, Geometric Properties of Banach Spaces and Nonlinear Iterations, Vol. 1965 of Lecture Notes in Mathematics, Springer London, London, 2009. doi:10.1007/978-1-84882-190-3.

[22] Y. Censor, A. Zenios, Parallel Optimization - Theory, Algorithms, and Applications, 1st Edition, Oxford University Press, New York, Oxford, 1997.

[23] B. D. Roberts, On the geometry of abstract vector spaces, Tohoku Mathematical Journal, First Series 39 (1927) (1934) 42–59.

[24] G. Birkhoff, Orthogonality in linear metric spaces, Duke Mathematical Journal 1 (1935) 169–172.

[25] R. C. James, Orthogonality in normed linear spaces, Duke Mathematical Journal 12 (2) (1945) 291–302. doi:10.1215/S0012-7094-45-01223-3.

[26] J. Muscat, Functional Analysis, 1st Edition, Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-06728-5.

[27] Y.-X. Yuan, J. Stoer, A subspace study on conjugate gradient algorithms, ZAMM - Journal of Applied Mathematics and Mechanics 75 (11) (1995) 69–77.

[28] N. Dunford, J. T. Schwartz, Linear Operators - Part I - General Theory, 1st Edition, Interscience Publishers, Inc., New York, 1957.

[29] E. Zeidler, Nonlinear functional analysis and its applications. 1. Fixed-point theorems, 1st Edition, Springer-Verlag Berlin Heidelberg New York, Berlin Heidelberg New York, 1986.

[30] S. F. Ashby, T. A. Manteuffel, P. E. Saylor, A Taxonomy for Conjugate Gradient Methods, SIAM Journal on Numerical Analysis 27 (6) (1990) 1542–1568. doi:10.1137/0727091.

[31] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.

[32] A. Chronopoulos, Z. Zlatev, Iterative methods for nonlinear operator equations, Applied Mathematics and Computation 51 (2-3) (1992) 167–180. doi:10.1016/0096-3003(92)90072-9.

[33] G. H. Golub, D. P. O'Leary, Some History of the Conjugate Gradient and Lanczos Methods, SIAM Review 31 (1) (1989) 50–102.

[34] C. Lanczos, Solution of systems of linear equations by minimized iterations, Journal Of Research Of The National Bureau Of Standards Section B Mathematics And Mathematical 49 (1) (1952) 33–53. doi:10.6028/jres.049.006.

[35] G. H. Golub, C. F. van Loan, Matrix Computations, The Johns Hopkins University Press, London, 1996.

[36] G. Guennebaud, B. Jacob, Others, Eigen v3 (2010).
URL http://eigen.tuxfamily.org

[37] I. Loris, On the performance of algorithms for the minimization of l1 - penalized functionals, Inverse Problems 25 (3) (2009) 035008. doi:10.1088/0266-5611/25/3/035008.

[38] F. Natterer, The Mathematics of Computerized Tomography, Wiley, Chichester, 1986.

[39] F. Natterer, F. Wübbeling, Mathematical Methods in Image Reconstruction, SIAM, Philadelphia, 2001.

[40] P. C. Hansen, Discrete Inverse Problems - Insight and Algorithms, 1st Edition, SIAM, 2010.

[41] N. Andrei, Another nonlinear conjugate gradient algorithm for unconstrained optimization, Optimization Methods and Software 24 (1) (2009) 89–104. doi:10.1080/10556780802393326.

[42] A. Nemirovskii, The regularizing properties of the adjoint gradient method in ill-posed problems, USSR Computational Mathematics and Mathematical Physics 26 (2) (1986) 7–16. doi:10.1016/0041-5553(86)90002-9.

[43] M. Hanke, Conjugate gradient type methods for ill-posed problems, Vol. 327 of Pitman Research Notes in Mathematics Series, Longman Scientific & Technical, Burnt Mill, Harlow, England, 1995.