# Eventual linear convergence of the Douglas-Rachford iteration for basis pursuit

Laurent Demanet*and Xiangxiong Zhang†

Massachusetts Institute of Technology, Department of Mathematics,
77 Massachusetts Avenue, Cambridge, MA 02139

December 2012, revised April 2013

**Abstract**

We provide a simple analysis of the Douglas-Rachford splitting algorithm in the context of $\ell^1$ minimization with linear constraints, and quantify the asymptotic linear convergence rate in terms of principal angles between relevant vector spaces. In the compressed sensing setting, we show how to bound this rate in terms of the restricted isometry constant. More general iterative schemes obtained by $\ell^2$-regularization and over-relaxation including the dual split Bregman method [24] are also treated. We make no attempt at characterizing the transient regime preceding the onset of linear convergence.

**Acknowledgments:** The authors are grateful to Jalal Fadili, Stanley Osher, Gabriel Peyré, Ming Yan, Yi Yang and Wotao Yin for discussions on modern methods of optimization that were very instructive to us. The authors are supported by the National Science Foundation and the Alfred P. Sloan Foundation.

**Keywords:** basis pursuit; Douglas-Rachford; asymptotic linear convergence rate

*laurent@math.mit.edu
†zhangxx@math.mit.edu

# 1 Introduction

## 1.1 Setup

In this paper we consider certain splitting algorithms for basis pursuit [7], the constrained optimization problem

$$\min \|x\|_1 \qquad \text{s.t.} \qquad Ax = b. \tag{1.1}$$

Throughout this paper we consider $A \in \mathbb{R}^{m\times n}$ with $m \leq n$, and we assume that $A$ has full row rank. We also assume that the solution $x^*$ of (1.1) is unique.

In particular, we treat splitting algorithms that naturally arise in the scope of minimization problems of the form

$$\min_x f(x) + g(x),$$

where $f$ and $g$ are convex, lower semi-continuous (but not otherwise smooth), and have simple resolvents

$$J_{\gamma F} = (\mathrm{I} + \gamma F)^{-1}, \qquad J_{\gamma G} = (\mathrm{I} + \gamma G)^{-1},$$

where $F = \partial f(x)$ and $G = \partial g(x)$ are the respective subdifferentials of $f$ and $g$ at $x$. In those terms, $x$ is a minimizer if and only if $0 \in F(x) + G(x)$. Resolvents are also often called proximal operators, as they obey $J_{\gamma F}(x) = \arg\min_z \gamma f(z) + \frac{1}{2}\|z - x\|^2$. In the case of basis pursuit, it is well known that

- $f(x) = \|x\|_1$ and $g(x) = \iota_{\{x:Ax=b\}}$, the indicator function equal to zero when $Ax = b$ and $+\infty$ otherwise;
- $J_{\gamma F}$ is soft-thresholding (shrinkage) by an amount $\gamma$,

$$J_{\gamma F}(x)_i = S_\gamma(x)_i = \operatorname{sgn}(x_i)\max\{|x_i| - \gamma, 0\};$$

- $J_{\gamma G}$ is projection onto the set $Ax = b$, namely

$$J_{\gamma G}(x) = \mathrm{P}(x) = x + A^+(b - Ax),$$

  with $A^+ = A^T(AA^T)^{-1}$ denoting the pseudo inverse.

The simplest splitting algorithm based on the resolvents is

$$x^{k+1} = J_{\gamma F} J_{\gamma G} x^k.$$

This iteration is successful in the special case when $f$ and $g$ are both indicators of convex sets, but does not otherwise generally enjoy good convergence properties. Instead, one is led to consider reflection operators $R_{\gamma F} = 2J_{\gamma F} - I$, $R_{\gamma G} = 2J_{\gamma G} - I$, and write the *Douglas-Rachford splitting* [22, 9]

$$\begin{cases} y^{k+1} = \frac{1}{2}(R_{\gamma F} R_{\gamma G} + I)y^k = J_{\gamma F} \circ (2J_{\gamma G} - I)y^k + (I - J_{\gamma G})y^k, \\ x^{k+1} = J_{\gamma G} y^{k+1}, \end{cases} \tag{1.2}$$

where $I$ is the identity. The operator $T_\gamma = \frac{1}{2}(R_{\gamma F} R_{\gamma G} + I)$ is *firmly non-expansive* regardless of $\gamma > 0$ [22]. Thus $y^k$ converges to one of its fixed points $y^*$. Moreover, $x^* = J_{\gamma G}(y^*)$ is one solution to $0 \in F(x) + G(x)$.

For general convex functions $f(x)$ and $g(x)$, the sublinear convergence rate $\mathcal{O}(1/k)$ of the algorithm (1.2) was proven for averages of iterates in [6, 16]. The firm non-expansiveness also implies $\|y^k - y^{k-1}\|^2 \leq \frac{1}{k}\|y^0 - y^*\|^2$, see Appendix A. Convergence questions for the Douglas-Rachford splitting were recently studied in the context of projections onto possibly nonconvex sets [1, 19] with potential applications to phase retrieval [2].

In the case of basis pursuit, we note that the Douglas-Rachford (DR) iteration takes the form

$$\begin{cases} y^{k+1} = S_\gamma(2x^k - y^k) + y^k - x^k, \\ x^{k+1} = y^{k+1} + A^+(b - Ay^{k+1}) \end{cases} . \tag{1.3}$$

For convenience, we use $\mathrm{R} = 2\mathrm{P} - \mathrm{I}$ to denote reflection about $Ax = b$, i.e., $\mathrm{R}(x) = x + 2A^+(b - Ax)$. It is easy to see that R is idempotent. Then $T_\gamma = S_\gamma \circ \mathrm{R} + \mathrm{I} - \mathrm{P}$.

## 1.2 Main result

In practice one often observes fast convergence for (1.3). For instance, see Figure 1.1 for an illustration of a typical error curve where the matrix $A$ is a $3 \times 40$ random matrix and $x^*$ has three nonzero components. Notice that the error $\|y^k - y^*\|$ is monotonically decreasing since the operator $T_\gamma$ is non-expansive. The same cannot be said of $\|x^k - x^*\|$. The iterations quickly settled into linear convergence of the $y^k$

In this example, the regime of linear convergence was reached quickly for the $y^k$. That may not in general be the case, particularly if $AA^T$ is ill-conditioned. Below, we provide the characterization of the error decay rate in the linear regime. To express the result, we need the following notations.

Assume that the unique solution $x^*$ of (1.1) has $r$ zero components. Let $e_i$ $(i = 1, \cdots, n)$ be the standard basis in $\mathbb{R}^n$. Denote the basis vectors corresponding to zero components in $x^*$ as $e_j$ $(j = i_1, \cdots, i_r)$. Let $B$ be the $r \times n$ selector of the zero components of $x^*$, i.e., $B = [e_{i_1}, \cdots, e_{i_r}]^T$. Let $\mathcal{N}(A) = \{x : Ax = 0\}$ denote the nullspace of $A$ and $\mathcal{R}(A^T) = \{x : x = A^T z, z \in \mathbb{R}^m\}$ denote the range of $A^T$.

Then, for the numerical example discussed earlier, the slope of $\log \|y^k - y^*\|$ as a function of $k$ is $\log(\cos\theta_1)$ for large $k$, where $\theta_1$ is the first principal angle between $\mathcal{N}(A)$ and $\mathcal{N}(B)$. See Definition 2.3 in Section 2.3 for principal angles between subspaces.

Our main result is that the rate of decay of the error is indeed $\cos\theta_1$ for a large class of situations that we call *standard*, in the sense of the following definition.

**Definition 1.1.** *Consider a basis pursuit problem $(b, A)$ with solution $x^*$. Consider $y^0$ an initial value for the Douglas-Rachford iteration, and $y^* = \lim_{k\to\infty} T_\gamma^k y^0$.*

*Consider the preimage of the soft thresholding of all vectors with the same signs as $x^*$:*

$$\mathcal{Q} = \{S_\gamma^{-1}(x) : \mathrm{sgn}(x) = \mathrm{sgn}(x^*)\} = Q_1 \otimes Q_2 \otimes \cdots Q_n,$$

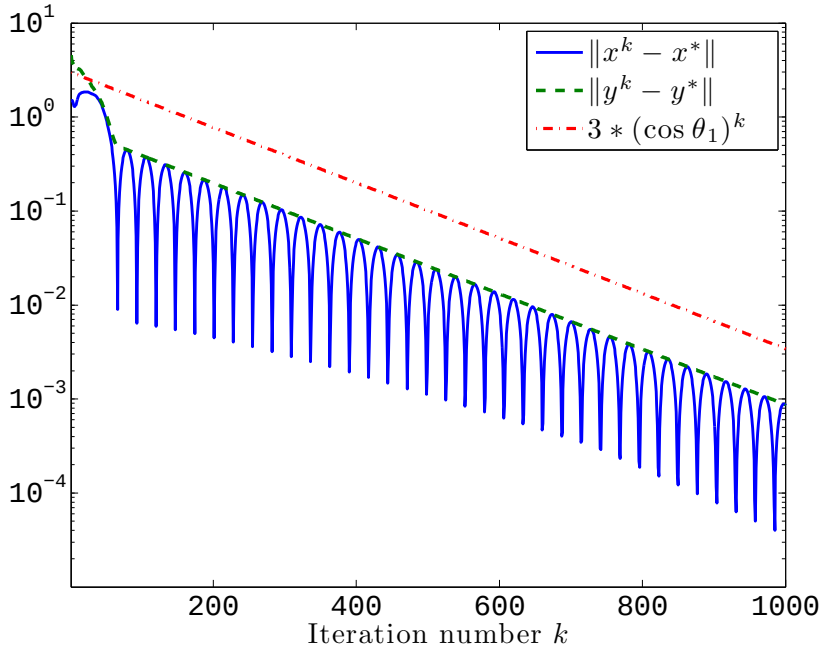



Figure 1.1: A typical error curve for Douglas-Rachford

*where*

$$Q_j = \begin{cases} (\gamma, +\infty), & \text{if } x_j^* > 0 \\ (-\infty, -\gamma), & \text{if } x_j^* < 0 \\ [-\gamma, \gamma], & \text{otherwise} \end{cases}.$$

*We call* $(b, A; y^0)$ *a* standard *problem for the Douglas-Rachford iteration if* $\mathrm{R}(y^*)$ *belongs to the interior of* $\mathcal{Q}$*, where* R *is the reflection operator defined earlier. In that case, we also say that the fixed point* $y^*$ *of* $T_\gamma$ *is an* interior fixed point*. Otherwise, we say that* $(b, A; y^0)$ *is nonstandard for the Douglas-Rachford iteration, and that* $y^*$ *is a* boundary fixed point*.*

**Theorem 1.2.** *Consider* $(b, A; y^0)$ *a standard problem for the Douglas-Rachford iteration, in the sense of the previous definition. Then the Douglas-Rachford iterates* $y^k$ *obey*

$$\|y^k - y^*\| \leq C\,(\cos\theta_1)^k\,,$$

*where* $C$ *may depend on* $b$*,* $A$ *and* $y^0$ *(but not on* $k$*), and* $\theta_1$ *is the leading principal angle between* $\mathcal{N}(A)$ *and* $\mathcal{N}(B)$*.*

The auxiliary variable $y^k$ in (1.3) converges linearly for sufficiently large $k$, thus $x^k$ is also bounded by a linearly convergent sequence since $\|x^k - x^*\| = \|\mathrm{P}(y^k) - \mathrm{P}(y^*)\| = \|\mathrm{P}(y^k - y^*)\| \leq \|y^k - y^*\|$.

Intuitively, convergence enters the linear regime when the support of the iterates essentially matches that of $x^*$. By essentially, we mean that there is some technical consideration (embodied in our definition of a "standard problem") that this match of supports is not a

fluke and will continue to hold for all iterates from $k$ and on. When this linear regime is reached, our analysis in the standard case hinges on the simple fact that $T_\gamma(y^k) - y^*$ is a linear transformation on $y^k - y^*$ with an eigenvalue of maximal modulus equal to $\cos\theta_1$.

In the nonstandard case ($y^*$ being a boundary fixed point), we furthermore show that the rate of convergence for $y^k$ is *generically* of the form $\cos\bar{\theta}_1$, where $0 < \bar{\theta}_1 \leq \theta_1$ is the leading principal angle between $\mathcal{N}(A)$ and $\mathcal{N}(\bar{B})$, with $\bar{B}$ a submatrix of $B$ depending on $y^*$. Nongeneric cases are not a priori excluded by our analysis, but have not been observed in our numerical tests. See Section 2.5 for a discussion of the different types of nonstandard cases.

### 1.3 Contributions

There is neither strong convexity nor Lipschitz continuity in the objective function of (1.1) even locally around $x^*$, but any $x^k$ with the same support as $x^*$ lies on a low-dimensional manifold, on which the objective function $\|x\|_1$ is smooth. Such property is characterized as *partial smoothness* [21]. In other words, it is not surprising that nonsmooth optimization algorithms for (1.1) converge linearly if $x^k$ has the correct support. For example, see [26] for the fast convergence of Bregman iterations.

The main contribution of this paper is the quantification of the asymptotic linear convergence rate for Douglas-Rachford splitting on (1.1). It is well-known that Douglas-Rachford on the dual problem is the same as the alternating direction method of multipliers (ADMM) [12], which is also equivalent to split Bregman method [15]. Thus the analysis in this paper also applies to ADMM on the dual problem of (1.1), i.e., the dual split Bregman method for basis pursuit [24].

### 1.4 Contents

Details and proof of the main result will be shown in Section 2. In Sections 3, we apply the same methodology to obtain the asymptotic convergence rates for Douglas-Rachford splitting on the $\ell^2$-regularized basis pursuit. Numerical experiments illustrating the theorems are shown. In Section 4, we discuss the dual split Bregman method and its practical relevance.

## 2 Douglas-Rachford for Basis Pursuit

### 2.1 Preliminaries

For any subspace $\mathcal{X}$ in $\mathbb{R}^n$, we use $\mathbb{P}_{\mathcal{X}}(z)$ to denote the orthogonal projection onto $\mathcal{X}$ of the point $z \in \mathbb{R}^n$.

As previously, we denote $F(x) = \partial\|x\|_1$, $G(x) = \partial\iota_{\{x:Ax=b\}}$, and the resolvents are $J_{\gamma F}(x) = S_\gamma(x)$ and $J_{\gamma G}(x) = \mathrm{P}(x) = x + A^+(b - Ax)$.

Let $N(x^*)$ denote the set of coordinate indices associated with the nonzero components of $x^*$, namely, $N(x^*) \cup \{i_1, \cdots, i_r\} = \{1, \cdots, n\}$. Recall the definition of $\mathcal{Q}$ in the previous section. Then for any $z \in \mathcal{Q}$, the soft thresholding operator can be written as $S_\gamma(z) = z - \gamma \sum_{j \in N(x^*)} \mathrm{sgn}(x_j^*)e_j - B^+Bz$.

**Lemma 2.1.** *The assumption that $x^*$ is the unique minimizer of (1.1) implies $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$.*

*Proof.* Suppose there exists a nonzero vector $z \in \mathcal{N}(A) \cap \mathcal{N}(B)$. For any $\varepsilon \in \mathbb{R}$ with small magnitude, we have $\operatorname{sgn}(x^* + \varepsilon z)^T = \operatorname{sgn}(x^*)^T$ and $A(x^* + \varepsilon z) = b$. For nonzero small $\varepsilon$, the uniqueness of the minimizer implies $\|x^*\|_1 < \|x^* + \varepsilon z\|_1 = \operatorname{sgn}(x^* + \varepsilon z)^T(x^* + \varepsilon z) = \operatorname{sgn}(x^*)^T(x^* + \varepsilon z) = \|x^*\|_1 + \varepsilon \operatorname{sgn}(x^*)^T z$. Thus $\operatorname{sgn}(x^*)^T z \neq 0$.

On the other hand, for the function $h(\varepsilon) = \|x^* + \varepsilon z\|_1 = \|x^*\|_1 + \varepsilon \operatorname{sgn}(x^*)^T z$ on a small neighborhood of $\varepsilon = 0$, the minimum of $h(\varepsilon)$ is $h(0)$, thus $\operatorname{sgn}(x^*)^T z = h'(0) = 0$. This contradicts with the fact that $\operatorname{sgn}(x^*)^T z \neq 0$. □

The sum of the dimensions of $\mathcal{N}(A)$ and $\mathcal{N}(B)$ should be no larger than $n$ since $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$. Thus, $n - m + n - r \leq n$ implies $m \geq n - r$.

$\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$ also implies the orthogonal complement of the subspace spanned by $\mathcal{N}(A)$ and $\mathcal{N}(B)$ is $\mathcal{R}(A^T) \cap \mathcal{R}(B^T)$. Therefore, the dimension of $\mathcal{R}(A^T) \cap \mathcal{R}(B^T)$ is $m + r - n$.

## 2.2 Characterization of the fixed points of $T_\gamma$

Since $\partial \iota_{\{x:Ax=b\}} = \mathcal{R}(A^T)$, the first order optimality condition for (1.1) reads $0 \in \partial\|x^*\|_1 + \mathcal{R}(A^T)$, thus $\partial\|x^*\|_1 \cap \mathcal{R}(A^T) \neq \varnothing$. Any such $\eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$ is called a dual certificate.

We have the following characterization of the fixed points of $T_\gamma$.

**Lemma 2.2.** *The set of the fixed points of $T_\gamma$ can be described as*

$$\{y^* : y^* = x^* - \gamma\eta, \eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)\}.$$

*Moreover, for any two fixed points $y_1^*$ and $y_2^*$, we have $y_1^* - y_2^*, \mathrm{R}(y_1^*) - \mathrm{R}(y_2^*) \in \mathcal{R}(A^T) \cap \mathcal{R}(B^T)$. Thus there is a unique fixed point $y^*$ if and only if $\mathcal{R}(A^T) \cap \mathcal{R}(B^T) = \{\mathbf{0}\}$.*

*Proof.* For any $\eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$, consider the vector $y^* = x^* - \gamma\eta$. Since $Ax^* = b$ and $A^+A\eta = \eta$ (implied by $\eta \in \mathcal{R}(A^T)$), we have $\mathrm{P}(y^*) = y^* + A^+(b - Ay^*) = x^* - \gamma\eta + A^+(b - Ax^* + A\gamma\eta) = x^* + A^+(b - Ax^*) = x^*$. Further, $\eta \in \partial\|x^*\|_1$ implies $S_\gamma(x^* + \gamma\eta) = x^*$. Thus $T_\gamma(y^*) = S_\gamma(2x^* - y^*) + y^* - x^* = S_\gamma(x^* + \gamma\eta) - x^* + y^* = y^*$.

Second, for any fixed point $y^*$ of the operator $T_\gamma$, let $\eta = (x^* - y^*)/\gamma$. Then

$$\mathrm{P}(y^*) = x^*, \quad \text{(see Theorem 5 in [9])} \tag{2.1}$$

implies $\eta = A^+A\eta$, thus $\eta \in \mathcal{R}(A^T)$. Further, $y^* = T_\gamma(y^*)$ implies $S_\gamma(x^* + \gamma\eta) = x^*$. We have $x^* = \arg\min_z \gamma\|z\|_1 + \frac{1}{2}\|z - (x^* + \gamma\eta)\|^2$, thus $\eta \in \partial\|x^*\|_1$.

Finally, let $y_1^*$ and $y_2^*$ be two fixed points. Then $y_1^* - y_2^* = -\gamma(\eta_1 - \eta_2)$ and $\mathrm{R}(y_1^*) - \mathrm{R}(y_2^*) = \gamma(\eta_1 - \eta_2)$ for some $\eta_1, \eta_2 \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$. Notice that $\eta_1, \eta_2 \in \partial\|x^*\|_1$ implies $\eta_1 - \eta_2 \in \mathcal{R}(B^T)$. So we get $y_1^* - y_2^*, \mathrm{R}(y_1^*) - \mathrm{R}(y_2^*) \in \mathcal{R}(A^T) \cap \mathcal{R}(B^T)$. □

With the assumption the matrix $A$ has full row rank, the following condition is sufficient [11] and necessary [27] to ensure existence of a unique solution $x^*$ to (1.1):

1. those columns of $A$ with respect to the support of $x^*$ are linearly independent.

2. there exists a dual certificate $\eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$ such that $\mathbb{P}_{\mathcal{N}(B)}(\eta) = \mathrm{sgn}(x^*)$ and $\|\mathbb{P}_{\mathcal{R}(B^T)}(\eta)\|_\infty < 1$.

Therefore, with assumption that there is a unique solution $x^*$ to (1.1), there always exists a dual certificate $\eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$ such that $\mathbb{P}_{\mathcal{N}(B)}(\eta) = \mathbb{P}_{\mathcal{N}(B)}(x^*)$ and $\|\mathbb{P}_{\mathcal{R}(B^T)}(\eta)\|_\infty < 1$. By Lemma 2.2, $y^* = x^* - \gamma\eta$ is a fixed point. And $\mathrm{R}(y^*)$ is in the interior of $\mathcal{Q}$ since $\mathrm{R}(y^*) = \mathrm{R}(x^* - \gamma\eta) = x^* + \gamma\eta$.

We call a fixed point $y^*$ an *interior fixed point* if $\mathrm{R}(y^*)$ is in the interior of the set $\mathcal{Q}$, or a *boundary fixed point* otherwise. A boundary fixed point exists only if $\mathcal{R}(A^T) \cap \mathcal{R}(B^T) \neq \{\mathbf{0}\}$.

**Definition 2.3.** *Let $\mathcal{U}$ and $\mathcal{V}$ be two subspaces of $\mathbb{R}^n$ with $dim(\mathcal{U}) = p \leq dim(\mathcal{V})$. The principal angles $\theta_k \in [0, \frac{\pi}{2}]$ $(k = 1, \cdots, p)$ between $\mathcal{U}$ and $\mathcal{V}$ are recursively defined by*

$$\cos\theta_k = \max_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} u^T v = u_k^T v_k, \quad \|u\| = \|v\| = 1,$$

$$u_j^T u = 0, \quad , u_j^T u = 0, \quad , j = 1, 2, \cdots, k-1.$$

*The vectors $(u_1, \cdots, u_p)$ and $(v_1, \cdots, v_p)$ are called principal vectors.*

**Lemma 2.4.** *Assume $y^*$ is a boundary fixed point and $\mathrm{R}(y^*)$ lies on a L-dimensional face of the set $\mathcal{Q}$. Namely, there are L coordinates $j_1, \cdots, j_L$ such that $|\mathrm{R}(y^*)_{j_l}| = \gamma$ $(l = 1, \cdots, L)$. Recall that $B = [e_{i_1}, \cdots, e_{i_r}]^T$, hence $\{j_1, \cdots, j_L\}$ is a subset of $\{i_1, \cdots, i_r\}$. Let $B_1$ denote the $(r-1) \times n$ matrix consisting of all row vectors of $B$ except $[e_{j_1}]^T$. Recursively define $B_l$ as the $(r-l) \times n$ matrix consisting of all row vectors of $B_{l-1}$ except $[e_{j_l}]^T$ for $l = 2, \cdots, L$. If there exists an index $l$ such that $\mathcal{R}(A^T) \cap \mathcal{R}(B_l^T) = \mathbf{0}$, let $M$ be the smallest such integer; otherwise, let $M = L$. Then $M \leq \dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B^T)\right]$, and the first principal angle between $\mathcal{N}(A)$ and $\mathcal{N}(B_l)$ $(l = 1, \cdots, M)$ is nonzero.*

*Proof.* Let $\mathcal{R}_l$ $(l = 1, \cdots, L)$ denote the one dimensional subspaces spanned by $e_{j_l}$, then $\mathcal{R}(B_{l-1}) = \mathcal{R}_l \oplus \mathcal{R}(B_l)$ and $\mathcal{N}(B_l) = \mathcal{R}_l \oplus \mathcal{N}(B_{l-1})$.

Let $z^*$ be an interior fixed point. Notice that $|\mathrm{R}(y^*)_{j_l}| = \gamma$ and $|\mathrm{R}(z^*)_{j_l}| < \gamma$ for each $l = 1, \cdots, L$, thus $\mathbb{P}_{\mathcal{R}_l}[\mathrm{R}(y^*) - \mathrm{R}(z^*)] = \mathrm{R}(y^*)_{j_l} - \mathrm{R}(z^*)_{j_l} \neq \mathbf{0}$. By Lemma 2.2 we have $\mathrm{R}(y^*) - \mathrm{R}(z^*) \in \mathcal{R}(A^T) \cap \mathcal{R}(B^T)$, therefore

$$\mathcal{R}_l \nsubseteq (\mathcal{R}(A^T) \cap \mathcal{R}(B^T))^\perp, \quad \forall l = 1, \cdots, L. \tag{2.2}$$

Since $\mathcal{R}(B^T) = \mathcal{R}(B_l^T) \oplus \mathcal{R}_1 \oplus \cdots \oplus \mathcal{R}_{l-1}$, with (2.2), we conclude that

$$\dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B_1^T)\right] \leq \dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B^T)\right] - 1.$$

Similarly, we have

$$\dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B_l^T)\right] \leq \dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B_{l-1}^T)\right] - 1, \quad l = 1, \cdots, M.$$

Therefore,

$$\dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B_l^T)\right] \leq \dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B^T)\right] - l, \quad \forall l = 1, \cdots, M, \tag{2.3}$$

thus $M \leq \dim\left[\mathcal{R}(A^T) \cap \mathcal{R}(B^T)\right]$.

Let $\mathcal{N}(A) \cup \mathcal{N}(B)$ denote the subspace spanned by $\mathcal{N}(A)$ and $\mathcal{N}(B)$. Since $\mathbb{R}^n = [\mathcal{R}(A^T) \cap \mathcal{R}(B^T)] \oplus [\mathcal{N}(A) \cup \mathcal{N}(B)] = [\mathcal{R}(A^T) \cap \mathcal{R}(B_l^T)] \oplus [\mathcal{N}(A) \cup \mathcal{N}(B_l)]$, by(2.3), we have $\dim[\mathcal{N}(A) \cup \mathcal{N}(B_l)] \geq \dim[\mathcal{N}(A) \cup \mathcal{N}(B)] + l = \dim[\mathcal{N}(A)] + \dim[\mathcal{N}(B)] + l = \dim[\mathcal{N}(A)] + \dim[\mathcal{N}(B_l)]$ for $(l = 1, \cdots, M)$. Therefore $\mathcal{N}(A) \cap \mathcal{N}(B_l) = \mathbf{0}$, and the first principal angle between $\mathcal{N}(A)$ and $\mathcal{N}(B_l)$ is nonzero. □

## 2.3 The characterization of the operator $T_\gamma$

**Lemma 2.5.** *For any $y$ satisfying $\mathrm{R}(y) \in \mathcal{Q}$ and any fixed point $y^*$, $T_\gamma(y) - T_\gamma(y^*) = [(I_n - B^+B)(I_n - A^+A) + B^+BA^+A](y - y^*)$ where $I_n$ denotes the $n \times n$ identity matrix.*

*Proof.* First, we have

$$\begin{aligned} T_\gamma(y) &= [S_\gamma \circ (2\mathrm{P} - \mathrm{I}) + \mathrm{I} - \mathrm{P}](y) = S_\gamma(\mathrm{R}(y)) + y - \mathrm{P}(y) \\ &= \mathrm{R}(y) - \gamma \sum_{j \in N(x^*)} e_j \operatorname{sgn}(x_j^*) - B^+B\mathrm{R}(y) + y - \mathrm{P}(y) \\ &= \mathrm{P}(y) - \gamma \sum_{j \in N(x^*)} e_j \operatorname{sgn}(x_j^*) - B^+B\mathrm{R}(y). \end{aligned}$$

The last step is due to the fact $\mathrm{R} = 2\mathrm{P} - \mathrm{I}$. The definition of fixed points and (2.1) imply

$$S_\gamma(\mathrm{R}(y^*)) = x^*, \tag{2.4}$$

thus $\mathrm{R}(y^*) \in \mathcal{Q}$. So we also have

$$T_\gamma(y^*) = \mathrm{P}(y^*) - \gamma \sum_{j \in N(x^*)} e_j \operatorname{sgn}(x_j^*) - B^+B\mathrm{R}(y^*).$$

Let $v = y - y^*$, then

$$\begin{aligned} T_\gamma(y) - T_\gamma(y^*) &= \mathrm{P}(y) - B^+B\mathrm{R}(y) - \left[\mathrm{P}(y^*) - B^+B\mathrm{R}(y^*)\right] \\ &= y + A^+(b - Ay) - B^+B(y + 2A^+(b - Ay)) \\ &\quad - \left[y^* + A^+(b - Ay^*) - B^+B(y^* + 2A^+(b - Ay^*))\right] \\ &= v - A^+Av - B^+Bv + 2B^+BA^+Av \\ &= [(I_n - B^+B)(I_n - A^+A) + B^+BA^+A]v. \end{aligned}$$

□

We now study the matrix

$$\mathbf{T} = (I_n - B^+B)(I_n - A^+A) + B^+BA^+A. \tag{2.5}$$

Let $A_0$ be a $n \times (n-m)$ matrix whose column vectors form an orthonormal basis of $\mathcal{N}(A)$ and $A_1$ be a $n \times m$ matrix whose column vectors form an orthonormal basis of $\mathcal{R}(A^T)$. Since $A^+A$ represents the projection to $\mathcal{R}(A^T)$ and so is $A_1A_1^T$, we have $A^+A = A_1A_1^T$. Similarly, $I_n - A^+A = A_0A_0^T$. Let $B_0$ and $B_1$ be similarly defined for $\mathcal{N}(B)$ and $\mathcal{R}(B^T)$. The matrix $\mathbf{T}$ can now be written as

$$\mathbf{T} = B_0B_0^TA_0A_0^T + B_1B_1^TA_1A_1^T.$$

It will be convenient to study the norm of the matrix $\mathbf{T}$ in terms of principal angles between subspaces.

Without loss of generality, we assume $n - r \leq n - m$. Let $\theta_i$ $(i = 1, \cdots, n - r)$ be the principal angles between the subspaces $\mathcal{N}(A)$ and $\mathcal{N}(B)$. Then the first principal angle

$\theta_1 > 0$ since $\mathcal{N}(A) \cap \mathcal{N}(B) = \mathbf{0}$. Let $\cos\Theta$ denote the $(n-r)\times(n-r)$ diagonal matrix with the diagonal entries $(\cos\theta_1, \cdots, \cos\theta_{(n-r)})$.

The singular value decomposition (SVD) of the $(n-r)\times(n-m)$ matrix $E_0 = B_0^T A_0$ is $E_0 = U_0 \cos\Theta V^T$ with $U_0^T U_0 = V^T V = I_{(n-r)}$, and the column vectors of $B_0 U_0$ and $A_0 V$ give the principal vectors, see Theorem 1 in [3].

By the definition of SVD, $V$ is a $(n-m)\times(n-r)$ matrix and its column vectors are orthonormalized. Let $V'$ be a $(n-m)\times(r-m)$ matrix whose column vectors are normalized and orthogonal to those of $V$. For the matrix $\widetilde{V} = (V, V')$, we have $I_{(n-m)} = \widetilde{V}\widetilde{V}^T$. For the matrix $E_1 = B_1^T A_0$, consider $E_1^T E_1 = A_0^T B_1 B_1^T A_0$. Since $B_0 B_0^T + B_1 B_1^T = I_n$, we have $E_1^T E_1 = A_0^T A_0 - A_0^T B_0 B_0^T A_0 = I_{(n-m)} - V\cos^2\Theta V^T = (V, V') \begin{pmatrix} \sin^2\Theta & 0 \\ 0 & I_{(r-m)} \end{pmatrix} (V, V')^T$, so the SVD of $E_1$ can be written as

$$B_1^T A_0 = E_1 = U_1 \begin{pmatrix} \sin\Theta & 0 \\ 0 & I_{(r-m)} \end{pmatrix} \widetilde{V}^T. \tag{2.6}$$

Notice that $A_0 = B_0 B_0^T A_0 + B_1 B_1^T A_0 = B_0 E_0 + B_1 E_1$, so we have

$$\begin{aligned} A_0 A_0^T &= (B_0, B_1) \begin{pmatrix} E_0 E_0^T & E_0 E_1^T \\ E_1 E_0^T & E_1 E_1^T \end{pmatrix} (B_0, B_1)^T \\ &= (B_0 U_0, B_1 U_1) \left(\begin{array}{c|cc} \cos^2\Theta & \cos\Theta\sin\Theta & 0 \\ \hline \cos\Theta\sin\Theta & \sin^2\Theta & 0 \\ 0 & 0 & I_{(r-m)} \end{array}\right) (B_0 U_0, B_1 U_1)^T. \end{aligned} \tag{2.7}$$

Let $\mathcal{C}$ denote the orthogonal complement of $\mathcal{R}(A^T) \cap \mathcal{R}(B^T)$ in the subspace $\mathcal{R}(B^T)$, namely, $\mathcal{R}(B^T) = [\mathcal{R}(A^T) \cap \mathcal{R}(B^T)] \oplus \mathcal{C}$. Then the dimension of $\mathcal{C}$ is $n-m$. Let $\widetilde{B}_0 = B_0 U_0$ and $\widetilde{B}_1 = B_1 U_1$, then the column vectors of $\widetilde{B}_0$ form an orthonormal basis of $\mathcal{N}(B)$. The column vectors of $\widetilde{B}_1$ are a family of orthonormal vectors in $\mathcal{R}(B^T)$. Moreover, the SVD (2.6) implies the columns of $\widetilde{B}_1$ and $A_0\widetilde{V}$ are principal vectors corresponding to angles $\{\frac{\pi}{2} - \theta_1, \cdots, \frac{\pi}{2} - \theta_{(n-r)}, 0, \cdots, 0\}$ between the two subspaces $\mathcal{R}(B^T)$ and $\mathcal{N}(A)$ , see [3]. And $\theta_1 > 0$ implies the largest angle between $\mathcal{R}(B^T)$ and $\mathcal{N}(A)$ is less than $\pi/2$, so none of the column vectors of $\widetilde{B}_1$ is orthogonal to $\mathcal{N}(A)$ thus all the column vectors of $\widetilde{B}_1$ are in the subspace $\mathcal{C}$. By counting the dimension of $\mathcal{C}$, we know that column vectors of $\widetilde{B}_1$ form an orthonormal basis of $\mathcal{C}$.

Let $\widetilde{B}_2$ be a $n\times(r+m-n)$ whose columns form an orthonormal basis of $\mathcal{R}(A^T)\cap\mathcal{R}(B^T)$, then we have

$$A_0 A_0^T = (\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2) \left(\begin{array}{c|c|c|c} \cos^2\Theta & \cos\Theta\sin\Theta & 0 & 0 \\ \hline \cos\Theta\sin\Theta & \sin^2\Theta & 0 & 0 \\ \hline 0 & 0 & I_{(r-m)} & 0 \\ \hline 0 & 0 & 0 & 0_{(r+m-n)} \end{array}\right) \begin{pmatrix} \widetilde{B}_0^T \\ \widetilde{B}_1^T \\ \widetilde{B}_2^T \end{pmatrix}.$$

Since $(\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2)$ is a unitary matrix and $A_1 A_1^T = I_n - A_0 A_0^T$, we also have

$$A_1 A_1^T = (\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2) \left(\begin{array}{c|c|c|c} \sin^2\Theta & -\cos\Theta\sin\Theta & 0 & 0 \\ \hline -\cos\Theta\sin\Theta & \cos^2\Theta & 0 & 0 \\ \hline 0 & 0 & 0_{(r-m)} & 0 \\ \hline 0 & 0 & 0 & I_{(r+m-n)} \end{array}\right) \begin{pmatrix} \widetilde{B}_0^T \\ \widetilde{B}_1^T \\ \widetilde{B}_2^T \end{pmatrix}. \tag{2.8}$$

Therefore, we get the decomposition

$$\begin{aligned}\mathbf{T} &= B_0B_0^TA_0A_0^T + B_1B_1^TA_1A_1^T \\ &= (\widetilde{B}_0,\widetilde{B}_1,\widetilde{B}_2)\left(\begin{array}{ccc|c} \cos^2\Theta & \cos\Theta\sin\Theta & 0 & 0 \\ -\cos\Theta\sin\Theta & \cos^2\Theta & 0 & 0 \\ 0 & 0 & 0_{(r-m)} & 0 \\ \hline 0 & 0 & 0 & I_{(r+m-n)} \end{array}\right)\left(\begin{array}{c}\widetilde{B}_0^T\\ \widetilde{B}_1^T\\ \widetilde{B}_2^T\end{array}\right). \quad (2.9)\end{aligned}$$

## 2.4 Standard cases: the interior fixed points

Assume the sequence $y^k$ will converge to an interior fixed point.

First, consider the simple case when $\mathcal{R}(A^T)\cap\mathcal{R}(B^T)=\{\mathbf{0}\}$, then $m+r=n$ and the fixed point is unique and interior. Let $\mathcal{B}_a(z)$ denote the ball centered at $z$ with radius $a$. Let $\varepsilon$ be the largest number such that $\mathcal{B}_\varepsilon(\mathrm{R}(y^*))\subseteq\mathcal{Q}$. Let $K$ be the smallest integer such that $y^K\in\mathcal{B}_\varepsilon(y^*)$ (thus $\mathrm{R}(y^K)\in\mathcal{B}_\varepsilon(\mathrm{R}(y^*))$). By nonexpansiveness of $T_\gamma$ and R, we get $\mathrm{R}(y^k)\in\mathcal{B}_\varepsilon(\mathrm{R}(y^*))$ for any $k\geq K$. By a recursive application of Lemma 2.5, we have

$$T_\gamma(y^k)-y^*=\mathbf{T}(T_\gamma(y^{k-1})-y^*)=\cdots=\mathbf{T}^{k-K}(y^K-y^*),\quad \forall k>K.$$

Now, (2.9) and $\mathcal{R}(A^T)\cap\mathcal{R}(B^T)=\{\mathbf{0}\}$ imply $\|\mathbf{T}\|_2=\cos\theta_1$. Notice that $\mathbf{T}$ is normal, so we have $\|\mathbf{T}^q\|_2=\|\mathbf{T}\|_2^q$ for any positive integer $q$. Thus we get the convergence rate for large $k$:

$$\|T_\gamma(y^k)-y^*\|_2\leq(\cos\theta_1)^{k-K}\|y^K-y^*\|_2,\quad\forall k>K. \quad (2.10)$$

If $\mathcal{R}(A^T)\cap\mathcal{R}(B^T)\neq\{\mathbf{0}\}$, then there are many fixed points by Lemma 2.2. Let $\mathcal{I}$ be the set of all interior fixed points. For $z^*\in\mathcal{I}$, let $\varepsilon(z^*)$ be the largest number such that $\mathcal{B}_{\varepsilon(z^*)}(\mathrm{R}(z^*))\subseteq\mathcal{Q}$.

If $y^K\in\bigcup_{z^*\in\mathcal{I}}\mathcal{B}_{\varepsilon(z^*)}(z^*)$ for some $K$, then consider the Euclidean projection of $y^K$ to $\mathcal{I}$, denoted by $y^*$. Then $\mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(B^T)}(y^K-y^*)=\mathbf{0}$ since $y_1^*-y_2^*\in\mathcal{R}(A^T)\cap\mathcal{R}(B^T)$ for any $y_1^*,y_2^*\in\mathcal{I}$. By (2.9), $\mathcal{R}(A^T)\cap\mathcal{R}(B^T)$ is the eigenspace of eigenvalue 1 for the matrix $\mathbf{T}$. So we have $\|\mathbf{T}(y^K-y^*)\|\leq\cos\theta_1\|y^K-y^*\|$, thus the error estimate (2.10) still holds.

The sequence $y^k$ may converge to a different fixed points for each initial value $y^0$; the fixed point $y^*$ is the projection of $y^K$ to $\mathcal{I}$. Here $K$ is the smallest integer such that $y^K\in\bigcup_{z^*\in\mathcal{I}}\mathcal{B}_{\varepsilon(z^*)}(\mathrm{R}(z^*))$.

**Theorem 2.6.** *For the algorithm (1.2) solving (1.1), if $y^k$ converges to an interior fixed point, then there exists an integer $K$ such that (2.10) holds.*

## 2.5 Nonstandard cases: the boundary fixed points

Suppose $y^k$ converges to a boundary fixed point $y^*$. With the same notations in Lemma 2.4, for simplicity, we only discuss the case $M=1$. More general cases can be discussed similarly. Without loss of generality, assume $j_1=1$ and $\mathrm{R}(y^*)_1=\gamma$. Then the set $\mathcal{Q}$ is equal to $Q_1\oplus Q_2\oplus\cdots\oplus Q_n$, with $Q_1=[-\gamma,\gamma]$. Consider another set $\mathcal{Q}_1=(\gamma,+\infty)\oplus Q_2\oplus\cdots\oplus Q_n$. Any neighborhood of $\mathrm{R}(y^*)$ intersects both $\mathcal{Q}$ and $\mathcal{Q}_1$.

There are three cases:

I. the sequence $\mathrm{R}(y^k)$ stays in $\mathcal{Q}$ if $k$ is large enough,

II. the sequence $\mathrm{R}(y^k)$ stays in $\mathcal{Q}_1$ if $k$ is large enough,

III. for any $K$, there exists $k_1, k_2 > K$ such that $\mathrm{R}(y^{k_1}) \in \mathcal{Q}$ and $\mathrm{R}(y^{k_2}) \in \mathcal{Q}_1$.

**Case I.** Assume $y^k$ converges to $y^*$ and $\mathrm{R}(y^k)$ stay in $\mathcal{Q}$ for any $k \geq K$. Then $\mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(B^T)}(y^K - y^*)$ must be zero. Otherwise, by (2.9), we have $\lim_{k\to\infty} y^k - y^* = \mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(B^T)}(y^K - y^*) \neq \mathbf{0}$. By (2.9), the eigenspace of $\mathbf{T}$ associated with the eigenvalue 1 is $\mathcal{R}(A^T) \cap \mathcal{R}(B^T)$, so (2.10) still holds.

**Case II** Assume $y^k$ converges to $y^*$ and $\mathrm{R}(y^k)$ stay in $\mathcal{Q}_1$ for any $k \geq K$. Let $\bar{B} = [e_{i_2}, \cdots, e_{i_r}]^T$. Following Lemma 2.5, for any $y$ satisfying $\mathrm{R}(y) \in \mathcal{Q}_1$, we have $T_\gamma(y) - T_\gamma(y^*) = [(I_n - \bar{B}^+\bar{B})(I_n - A^+A) + \bar{B}^+\bar{B}A^+A](y - y^*)$.

Without loss of generality, assume $n - r + 1 \leq n - m$. Consider the $(n - r + 1)$ principal angles between $\mathcal{N}(A)$ and $\mathcal{N}(\bar{B})$ denoted by $(\bar{\theta}_1, \cdots, \bar{\theta}_{(n-r+1)})$. Let $\Theta_1$ denote the diagonal matrix with diagonal entries $(\bar{\theta}_1, \cdots, \bar{\theta}_{(n-r+1)})$. Then the matrix $\bar{\mathbf{T}} = (I_n - \bar{B}^+\bar{B})(I_n - A^+A) + \bar{B}^+\bar{B}A^+A$ can be written as

$$\bar{\mathbf{T}} = (\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2) \left(\begin{array}{c|c|c|c} \cos^2\Theta_1 & \cos\Theta_1 \sin\Theta_1 & 0 & 0 \\ \hline -\cos\Theta_1 \sin\Theta_1 & \cos^2\Theta_1 & 0 & 0 \\ 0 & 0 & 0_{(r-m-1)} & 0 \\ \hline 0 & 0 & 0 & I_{(r+m-n-1)} \end{array}\right) \begin{pmatrix} \widetilde{B}_0^T \\ \widetilde{B}_1^T \\ \widetilde{B}_2^T \end{pmatrix}, \tag{2.11}$$

where $(\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2)$ are redefined accordingly.

By Lemma 2.4, $\bar{\theta}_1 > 0$. Following the first case, we have $\mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(\bar{B}^T)}(y^K - y^*) = \mathbf{0}$. So

$$\|T_\gamma(y^k) - y^*\|_2 \leq (\cos\bar{\theta}_1)^{k-K}\|y^K - y^*\|_2, \quad \forall k > K.$$

Convergence is slower than previously, as $\bar{\theta}_1 \leq \theta_1$.

**Case III** Assume $y^k$ converges to $y^*$ and $\mathrm{R}(y^k)$ stay in $\mathcal{Q} \cup \mathcal{Q}_1$ for any $k \geq K$. Then $\mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(\bar{B}^T)}(y^K - y^*) = \mathbf{0}$. And for $y^k \in \mathcal{Q}_1$ we have $\|T_\gamma(y^k) - y^*\|_2 \leq (\cos\bar{\theta}_1)\|y^k - y^*\|_2$. Let $\mathcal{D}$ be the orthogonal complement of $\mathcal{R}(A^T) \cap \mathcal{R}(\bar{B}^T)$ in $\mathcal{R}(A^T) \cap \mathcal{R}(B^T)$, namely $\mathcal{R}(A^T) \cap \mathcal{R}(B^T) = \mathcal{R}(A^T) \cap \mathcal{R}(\bar{B}^T) \oplus \mathcal{D}$. For $y^k \in \mathcal{Q}$, we have $\|\mathbb{P}_{\mathcal{D}^\perp}(T_\gamma(y^k) - y^*)\|_2 \leq \cos\theta_1\|\mathbb{P}_{\mathcal{D}^\perp}(y^k - y^*)\|_2$ and $\mathbb{P}_{\mathcal{D}}(T_\gamma(y^k) - y^*) = \mathbb{P}_{\mathcal{D}}(y^k - y^*)$.

For the Case III, which we refer to as nongeneric cases, no convergence results like $\|T_\gamma(y^k) - y^*\|_2 \leq (\cos\bar{\theta}_1)\|y^k - y^*\|_2$ can be established since $\mathbb{P}_{\mathcal{D}}(T_\gamma(y^k) - y^*) = \mathbb{P}_{\mathcal{D}}(y^k - y^*)$ whenever $\mathrm{R}(y^k) \in \mathcal{Q}$. Even though it seems hard to exclude Case III from the analysis, it has not been observed in our numerical tests.

## 2.6 Generalized Douglas-Rachford

The following generalized Douglas-Rachford splitting was proposed in [9],

$$\begin{cases} y^{k+1} = y^k + \lambda_k \left[S_\gamma(2x^k - y^k) - x^k\right] \\ x^{k+1} = y^{k+1} + A^+(b - Ay^{k+1}) \end{cases}, \tag{2.12}$$

where $\lambda_k \in (0, 2)$.

Let $T_\gamma^\lambda = \mathrm{I} + \lambda\,[S_\gamma \circ (2\mathrm{P} - \mathrm{I}) - \mathrm{P}]$. Then any fixed point $y^*$ of $T_\gamma^\lambda$ satisfies $\mathrm{P}(y^*) = x^*$, [8]. So the fixed points set of $T_\gamma^\lambda$ is the same as the fixed points set of $T_\gamma$. Moreover, for any $y$ satisfying $\mathrm{R}(y) \in \mathcal{Q}$ and any fixed point $y^*$, $T_\gamma^\lambda(y) - T_\gamma^\lambda(y^*) = [I_n + \lambda(I_n - B^+B)(I_n - 2A^+A) - \lambda(I_n - A^+A)](y - y^*)$.

To find the asymptotic convergence rate of (2.12), it suffices to consider the matrix $\mathbf{T}_\lambda = I_n + \lambda(I_n - B^+B)(I_n - 2A^+A) - \lambda(I_n - A^+A) = (1-\lambda)I_n + \lambda\mathbf{T}$. By (2.9), we have

$$\mathbf{T}_\lambda = \widetilde{B}\left(\begin{array}{c|ccc|c} \cos^2\Theta + (1-\lambda)\sin^2\Theta & \lambda\cos\Theta\sin\Theta & 0 & 0 \\ \hline -\lambda\cos\Theta\sin\Theta & \cos^2\Theta + (1-\lambda)\sin^2\Theta & 0 & 0 \\ 0 & 0 & (1-\lambda)I_{(r-m)} & 0 \\ \hline 0 & 0 & 0 & I_{(r+m-n)} \end{array}\right)\widetilde{B}^T,$$

where $\widetilde{B} = (\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2)$.

Notice that $\mathbf{T}_\lambda$ is a normal matrix. By the discussion in Section 2, if $y^k$ in the iteration of (2.12) converges to an interior fixed point, the asymptotic convergence rate will be governed by the matrix

$$\mathbf{M}_\lambda = \left(\begin{array}{c|cc} \cos^2\Theta + (1-\lambda)\sin^2\Theta & \lambda\cos\Theta\sin\Theta & 0 \\ \hline -\lambda\cos\Theta\sin\Theta & \cos^2\Theta + (1-\lambda)\sin^2\Theta & 0 \\ 0 & 0 & (1-\lambda)I_{(r-m)} \end{array}\right).$$

Note that $\|\mathbf{M}_\lambda\| = \sqrt{\lambda(2-\lambda)\cos^2\theta_1 + (1-\lambda)^2} \geq \cos\theta_1$ for any $\lambda \in (0,2)$. Therefore, the asymptotic convergence rate of (2.12) is always slower than (1.3) if $\lambda \neq 1$. But this does not mean (1.3) is more efficient than (2.12) for $x^k$ to reach a given accuracy.

## 2.7 Relation to the Restricted Isometry Property

Let $A$ be a $m \times n$ random matrix and each column of $A$ is normalized, i.e., $\sum_i A_{ij}^2 = 1$ for each $j$. The Restricted Isometry Property (RIP) introduced in [5] is as follows.

**Definition 2.7.** *For each integer $s = 1, 2, \cdots$, the restricted isometry constants $\delta_s$ of $A$ is the smallest number such that*

$$(1 - \delta_s)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta_s)\|x\|^2, \tag{2.13}$$

*holds for all vectors $x$ with at most $s$ nonzero entries.*

In particular, any vector with the same support as $x^*$ can be denoted as $(I_n - B^+B)x$ for some $x \in \mathbb{R}^n$. The RIP (2.13) with $s = n - r$ implies

$$(1 - \delta_{(n-r)})\|(I_n - B^+B)x\|^2 \leq \|A(I_n - B^+B)x\|^2 \leq (1 + \delta_{(n-r)})\|(I_n - B^+B)x\|^2, \quad \forall x \in \mathbb{R}^n.$$

Let $d$ denote the smallest eigenvalue of $(AA^T)^{-1}$. Then $d > 0$ since we assume $A$ has full row rank. For any vector $y$, we have

$$\|A^+Ay\|^2 = y^TA^T[(AA^T)^{-1}]^TAA^T(AA^T)^{-1}Ay = y^TA^T[(AA^T)^{-1}]^TAy \geq d\|Ay\|^2,$$

where the last step is due to the Courant–Fischer–Weyl min-max principle.

Therefore, we get

$$\|A^+A(I_n - B^+B)x\|^2 \geq d\|A(I_n - B^+B)x\|^2 \geq d(1-\delta_{(n-r)})\|(I_n - B^+B)x\|^2, \quad \forall x \in \mathbb{R}^n, \tag{2.14}$$

We will show that (2.14) gives a lower bound of the first principal angle $\theta_1$ between two subspaces $\mathcal{N}(A)$ and $\mathcal{N}(B)$. Notice that (2.8) implies

$$A^+A(I_n - B^+B) = A_1A_1^TB_0B_0^T = (B_0U_0, B_1U_1)\left(\begin{array}{c|c} \sin^2\Theta & 0 \\ \hline -\cos\Theta\sin\Theta & 0 \\ 0 & 0 \end{array}\right)(B_0U_0, B_1U_1)^T,$$

by which we have $\|A^+A(I_n - B^+B)x\|^2 = x^T(B_0U_0, B_1U_1)\left(\begin{array}{c|c} \sin^2\Theta & 0 \\ \hline 0 & 0 \end{array}\right)(B_0U_0, B_1U_1)^Tx$.

Let $z = (B_0U_0, B_1U_1)^Tx$. Since $I_n - B^+B = (B_0U_0, B_1U_1)\left(\begin{array}{c|c} I_{(n-r)} & 0 \\ \hline 0 & 0 \end{array}\right)(B_0U_0, B_1U_1)^T$, (2.14) is equivalent to

$$z^T\left(\begin{array}{c|c} \sin^2\Theta & 0 \\ \hline 0 & 0 \end{array}\right)z \geq d(1-\delta_{(n-r)})z^T\left(\begin{array}{c|c} I_{n-r} & 0 \\ \hline 0 & 0 \end{array}\right)z, \quad \forall z \in \mathbb{R}^n,$$

which implies $\sin^2\theta_1 \geq d(1-\delta_{(n-r)})$ by the Courant–Fischer–Weyl min-max principle. So the RIP constant gives us

$$\cos\theta_1 \leq \sqrt{1 - d(1-\delta_{(n-r)})}.$$

## 2.8 Numerical examples

We consider several examples for (1.3). In all the examples, $y^0 = \mathbf{0}$ unless specified otherwise. For examples in this subsection, the angles between the null spaces can be computed by singular value decomposition (SVD) of $A_0^TB_0$ [3].

**Example 1** The matrix $A$ is a $3 \times 40$ random matrix with standard normal distribution and $x^*$ has three nonzero components. By counting dimensions, we know that $\mathcal{R}(A^T) \cap \mathcal{R}(B^T) = \{\mathbf{0}\}$. Therefore there is only one fixed point. See Figure 1.1 for the error curve of $x^k$ and $y^k$ with $\gamma = 1$. Obviously, the error $\|x^k - x^*\|$ is not monotonically decreasing but $\|y^k - y^*\|$ is since the operator $T_\gamma$ is non-expansive. And the slope of $\log\|y^k - y^*\|$ is exactly $\log(\cos\theta_1) = \log(0.9932)$ for large $k$.

**Example 2** The matrix $A$ is a $10 \times 1000$ random matrix with standard normal distribution and $x^*$ has ten nonzero components. Thus there is only one fixed point. See Figure 2.1 for the error curve of $y^k$ with $\gamma = 0.1, 1, 10$. We take $y^*$ as the result of (1.3) after $8 \times 10^4$ iterations. The slopes of $\log\|y^k - y^*\|$ for different $\gamma$ are exactly $\log(\cos\theta_1) = \log(0.9995)$ for large $k$.

**Example 3** The matrix $A$ is a $18 \times 100$ submatrix of a $100 \times 100$ Fourier matrix and $x^*$ has two nonzero components. There are interior and boundary fixed points. In this example, we

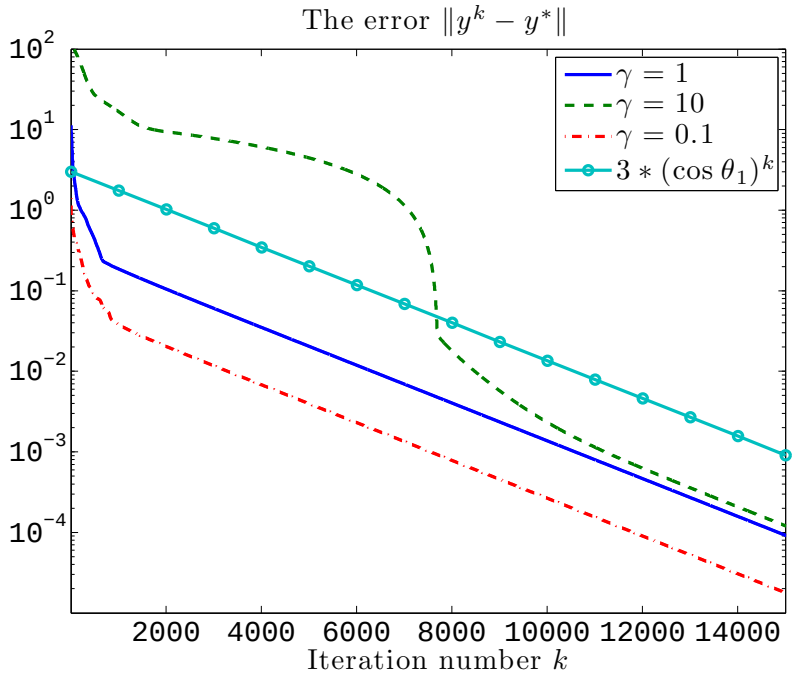



Figure 2.1: Example 2: for an interior fixed point, the asymptotic rate remains the same for different soft-thresholding parameter $\gamma$. The slope of the straight line is $\log(\cos\theta_1)$.

fix $\gamma = 1$ and test (1.3) with random $y^0$ for six times. See Figure 2.2 for the error curve of $x^k$. In Figure 2.2, in four tests, $y^k$ converges to an interior fix point, thus the convergence rate for large $k$ is governed by $\cos\theta_1 = 0.9163$. In the second and third tests, $y^k$ converges to different boundary fixed points[1] thus convergence rates are slower than $\cos\theta_1$. Nonetheless, the rate for large $k$ is still linear.

**Example 4** The matrix $A$ is a $5 \times 40$ random matrix with standard normal distribution and $x^*$ has three nonzero components. See Figure 2.3 for the comparison of (1.3) and (2.12) with $\gamma = 1$.

# 3 Douglas-Rachford for the $\ell^2$ regularized Basis Pursuit

Consider the regularized problem

$$\min_x \{\|x\|_1 + \frac{1}{2\alpha}\|x\|^2 : Ax = b\}. \tag{3.1}$$

It is proven in [25] that there exists a $\alpha_\infty$ such that the solution of (3.1) with $\alpha \geq \alpha_\infty$ is the solution of (1.1). See [20] for more discussion of $\alpha_\infty$. For the rest of this section, we assume $\alpha$ is taken large enough so that $\alpha \geq \alpha_\infty$.

[1] At least, numerically so in double precision.

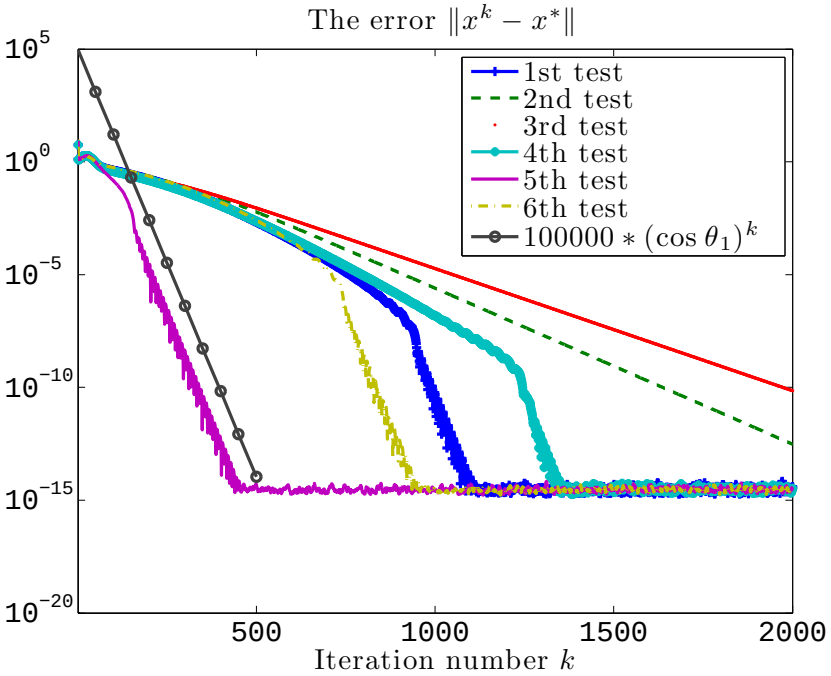



Figure 2.2: Example 3: fixed $\gamma = 1$ with random $y^0$.

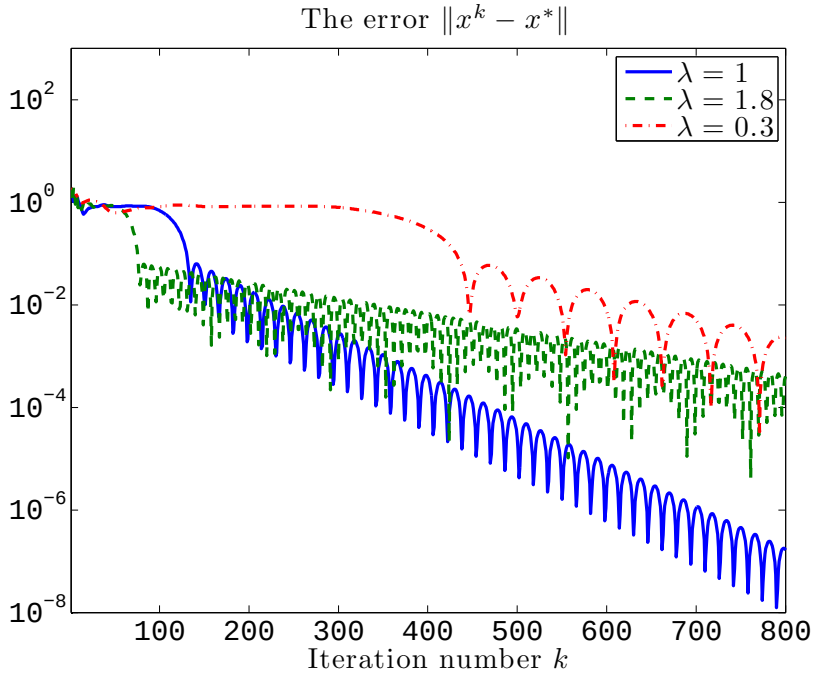



Figure 2.3: Example 4: The Generalized Douglas-Rachford (2.12) with different $\lambda$ and fixed $\gamma = 1$. The asymptotic convergence rate of (1.3) ($\lambda = 1$) is the fastest.

It is equivalent to solve the problem $\min_x \|x\|_1 + \iota_{\{x:Ax=b\}} + \frac{1}{2\alpha}\|x\|^2$. To use the Douglas-Rachford splitting, in particular, we can choose $f(x) = \|x\|_1 + \frac{1}{2\alpha}\|x\|^2$ and $g(x) = \iota_{\{x:Ax=b\}}$. The resolvent of $F = \partial f$ is $J_{\gamma F}(x) = \arg\min_z \gamma\|z\|_1 + \frac{\gamma}{2\alpha}\|z\|^2 + \frac{1}{2}\|z-x\|^2 = \arg\min_z \frac{\alpha\gamma}{\alpha+\gamma}\|z\|_1 + \frac{1}{2}\|z - \frac{\alpha}{\alpha+\gamma}x\|^2$. Therefore, $J_{\gamma F}(x) = \frac{\alpha}{\alpha+\gamma}S_\gamma(x)$. The Douglas-Rachford splitting (1.2) for (1.1) reads

$$\begin{cases} y^{k+1} = \frac{\alpha}{\alpha+\gamma}S_\gamma(2x^k - y^k) + y^k - x^k \\ x^{k+1} = y^{k+1} + A^+(b - Ay^{k+1}) \end{cases}. \tag{3.2}$$

Since $\|x\|_1 + \frac{1}{2\alpha}\|x\|^2$ is a strongly convex function, (3.1) always has a unique minimizer $x^*$ as long as $\{x : Ax = b\}$ is nonempty. The first order optimality condition $\mathbf{0} \in \partial F(x^*) + \partial G(x^*)$ implies the dual certificate set $(\partial\|x^*\|_1 + \frac{1}{\alpha}x^*) \cap \mathcal{R}(A^T)$ is nonempty. Let $T_\gamma^\alpha = \frac{\alpha}{\alpha+\gamma}S_\gamma \circ (2\mathrm{P} - \mathrm{I}) + \mathrm{I} - \mathrm{P}$.

**Lemma 3.1.** *The set of the fixed points of $T_\gamma^\alpha$ can be described as*

$$\left\{ y^* : y^* = x^* - \gamma\eta, \eta \in \left(\partial\|x^*\|_1 + \frac{1}{\alpha}x^*\right) \cap \mathcal{R}(A^T) \right\}.$$

The proof is similar to the one of Lemma 2.2. We also have

**Lemma 3.2.** *For any $y$ satisfying $\frac{\alpha}{\alpha+\gamma}\mathrm{R}(y) \in \mathcal{Q}$ and any fixed point $y^*$, $T_\gamma^\alpha(y) - T_\gamma^\alpha(y^*) = [c(I_n - B^+B)(I_n - A^+A) + cB^+BA^+A + (1-c)A^+A](y - y^*)$ where $c = \frac{\alpha}{\alpha+\gamma}$.*

*Proof.* First, we have

$$\begin{aligned} T_\gamma^\alpha(y) &= [cS_\gamma \circ (2\mathrm{P} - \mathrm{I}) + \mathrm{I} - \mathrm{P}](y) = cS_\gamma(\mathrm{R}(y)) + y - \mathrm{P}(y) \\ &= c\left[\mathrm{R}(y) - \gamma\sum_{j\in N(x^*)} e_j \operatorname{sgn}(x_j^*) - B^+B\mathrm{R}(y)\right] + y - \mathrm{P}(y) \end{aligned}$$

Similarly we also have

$$T_\gamma^\alpha(y^*) = c\left[\mathrm{R}(y^*) - \gamma\sum_{j\in N(x^*)} e_j \operatorname{sgn}(x_j^*) - B^+B\mathrm{R}(y^*)\right] + y^* - \mathrm{P}(y^*).$$

Let $v = y - y^*$, then

$$\begin{aligned} T_\gamma^\alpha(y) - T_\gamma^\alpha(y^*) &= c\left[\mathrm{R}(y) - B^+B\mathrm{R}(y)\right] + y - \mathrm{P}(y) \\ &\quad -c\left[\mathrm{R}(y^*) - B^+B\mathrm{R}(y^*)\right] - (y^* - \mathrm{P}(y^*)) \\ &= c[I_n - 2A^+A - B^+B + 2B^+BA^+A]v + A^+Av \\ &= [c(I_n - B^+B)(I_n - A^+A) + cB^+BA^+A + (1-c)A^+A]v. \end{aligned}$$

□

Consider the matrix $\mathbf{T}(c) = c(I_n - B^+B)(I_n - A^+A) + cB^+BA^+A + (1-c)A^+A$ with $c = \frac{\alpha}{\alpha+\gamma}$. Then $\mathbf{T}(c) = c\mathbf{T} + (1-c)A^+A$ where $\mathbf{T} = (I_n - B^+B)(I_n - A^+A) + B^+BA^+A$.

By (2.8) and (2.9), we have

$$\mathbf{T}(c) = (\widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2) \left(\begin{array}{c|c|c|c} (1-c)\sin^2\Theta + c\cos^2\Theta & (2c-1)\cos\Theta\sin\Theta & 0 & 0 \\ \hline -\cos\Theta\sin\Theta & \cos^2\Theta & 0 & 0 \\ \hline 0 & 0 & 0_{(r-m)} & 0 \\ \hline 0 & 0 & 0 & I_{(r+m-n)} \end{array}\right) \left(\begin{array}{c} \widetilde{B}_0^T \\ \widetilde{B}_1^T \\ \widetilde{B}_2^T \end{array}\right). \tag{3.3}$$

Following the proof in [27], it is straightforward to show there exists a dual certificate $\eta \in (\partial\|x^*\|_1 + \frac{1}{\alpha}x^*) \cap \mathcal{R}(A^T)$ such that $\mathbb{P}_{\mathcal{N}(B)}(\eta) = \mathbb{P}_{\mathcal{N}(B)}(x^*)$ and $\|\mathbb{P}_{\mathcal{R}(B^T)}(\eta)\|_\infty < 1$. So there is at least one interior fixed point. Following Lemma 2.2, there is only one fixed point $y^*$ if and only if $\mathcal{R}(A^T) \cap \mathcal{R}(B^T) = \{\mathbf{0}\}$.

For simplicity, we only discuss the interior fixed point case. The boundary fixed point case is similar to the previous discussion.

Assume $y^k$ converges to an interior fixed point $y^*$. Let $\varepsilon$ be the largest number such that $\mathcal{B}_\varepsilon(\mathrm{R}(y^*)) \subseteq S$. Let $K$ be the smallest integer such that $y^K \in \mathcal{B}_\varepsilon(y^*)$ (thus $\mathrm{R}(y^K) \in \mathcal{B}_\varepsilon(\mathrm{R}(y^*))$). By nonexpansiveness of $T_\gamma^\alpha$ and R, we get $\mathrm{R}(y^k) \in \mathcal{B}_\varepsilon(\mathrm{R}(y^*))$ for any $k \geq K$. So we have

$$\|y^k - y^*\| = \|\mathbf{T}(c)(y^k - y^*)\| = \cdots = \|\mathbf{T}(c)^{k-K}(y^K - y^*)\| \leq \|\mathbf{T}(c)^{k-K}\|\|(y^K - y^*)\|, \quad \forall k > K.$$

Notice that $\mathbf{T}(c)$ is a nonnormal matrix, so $\|\mathbf{T}(c)^q\|$ is much less than $\|\mathbf{T}(c)\|^q$ for large $q$. Thus the asymptotic convergence rate is governed by $\lim_{q\to\infty} \sqrt[q]{\|\mathbf{T}(c)^q\|}$, which is equal to the norm of the eigenvalues of $\mathbf{T}(c)$ with the largest magnitude.

It suffices to study the matrix $\mathbf{M}(c) = \left(\begin{array}{c|c} (1-c)\sin^2\Theta + c\cos^2\Theta & (2c-1)\cos\Theta\sin\Theta \\ \hline -\cos\Theta\sin\Theta & \cos^2\Theta \end{array}\right)$ because $\mathbb{P}_{\mathcal{R}(A^T)\cap\mathcal{R}(B^T)}(y^K - y^*) = \mathbf{0}$ (otherwise $y^k$ cannot converge to $y^*$).

Notice that $\det(\mathbf{M}(c) - \lambda\mathrm{I}) = \prod_{i=1}^{n-r} [\lambda^2 - (c\cos(2\theta_i) + 1)\lambda + c\cos^2\theta_i]$. Let $\lambda(\theta, c)$ denote the solution with the largest magnitude for the quadratic equation $\lambda^2 - (c\cos(2\theta)+1)\lambda + c\cos^2\theta$, with discriminant $\Delta = \cos^2(2\theta)c^2 - 2c + 1$.

The two solutions of $\Delta = 0$ are $[1 \pm \sin(2\theta)]/\cos^2(2\theta)$. Notice that $[1+\sin(2\theta)]/\cos^2(2\theta) \geq 1$ for $\theta \in [0, \pi/2]$ and $c \in (0,1)$, we have

$$|\lambda(\theta, c)| = \begin{cases} \sqrt{c}\cos\theta, & \text{if } c \geq \frac{1-\sin(2\theta)}{\cos^2(2\theta)} = \frac{1}{(\cos\theta+\sin\theta)^2} \\ \frac{1}{2}\left(c\cos(2\theta) + 1 + \sqrt{\cos^2(2\theta)c^2 - 2c + 1}\right) & \text{if } c \leq \frac{1}{(\cos\theta+\sin\theta)^2} \end{cases}. \tag{3.4}$$

It is straightforward to check that $|\lambda(\theta, c)|$ is monotonically decreasing with respect to $\theta$. Therefore, the asymptotic convergence rate is equal to $|\lambda(\theta_1, c)|$.

Let $c^* = \frac{1}{(\cos\theta_1+\sin\theta_1)^2}$ which is equal to $\arg\min_c |\lambda(\theta_1, c)|$. Let $c^\sharp = \frac{1}{1+2\cos\theta_1}$ which is the solution to $|\lambda(\theta_1, c)| = \cos\theta_1$. See Figure 3.1. Then for any $c \in (c^\sharp, 1)$, we have $|\lambda(\theta_1, c)| < \cos\theta_1$. Namely, the asymptotic convergence rate of (3.2) is faster than (1.3) if $\frac{\alpha}{\alpha+\gamma} \in (c^\sharp, 1)$. The best asymptotic convergence rate that (3.2) can achieve is $|\lambda(\theta_1, c^*)| = \sqrt{c^*}\cos\theta_1 = \frac{\cos\theta_1}{\cos\theta_1+\sin\theta_1} = \frac{1}{1+\tan\theta_1}$ when $\frac{\alpha}{\alpha+\gamma} = c^*$.

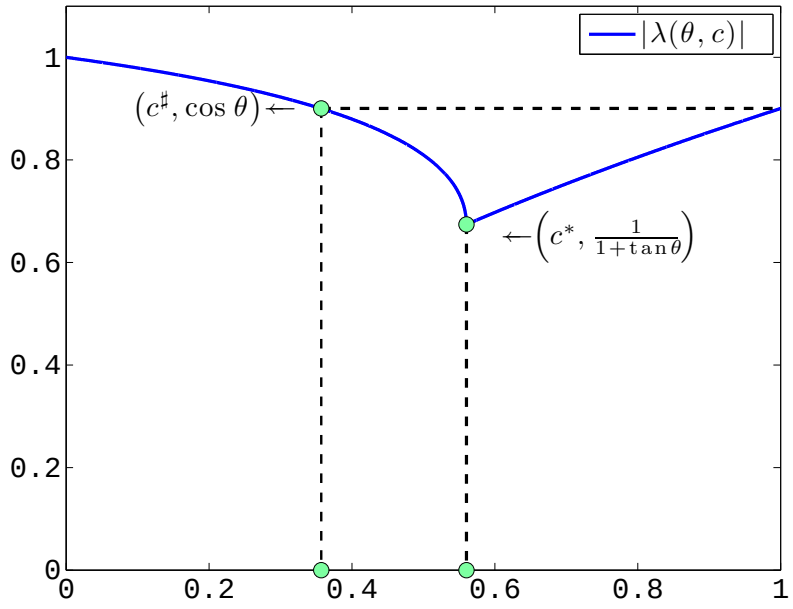



Figure 3.1: An illustration of (3.4) for a fixed $\theta$.

**Example 5** The matrix $A$ is a $40 \times 1000$ random matrix with standard normal distribution and $x^*$ has two nonzero components. This is a typical compressive sensing problem. We test the algorithm (3.2), which is the same as the dual split Bregman method in [24]. See Section 4.3 for the equivalence. See Figure 3.2 for the error curve of $x^k$. The best choice of the parameter according to Figure 3.1 should be $\alpha/(\alpha+\gamma) = c^*$ which is $c^* = 0.756$ for this example. Here $c^*$ indeed gives the best asymptotic rate $\frac{1}{1+\tan\theta_1}$ but $c^*$ is not necessarily the most efficient choice, as we can see in the figure.

# 4 Dual interpretation

## 4.1 Chambolle and Pock's primal dual algorithm

The algorithm (1.2) is equivalent to a special case of Chambolle and Pock's primal-dual algorithm [6]. Let $w^{k+1} = (x^k - y^{k+1})/\gamma$, then (1.2) with $F = \partial f$ and $G = \partial g$ is equivalent to

$$\begin{cases} w^{k+1} &= (\mathrm{I} + \frac{1}{\gamma}\partial f^*)^{-1}(w^k + \frac{1}{\gamma}(2x^k - x^{k-1})) \\ x^{k+1} &= (\mathrm{I} + \gamma\partial g)^{-1}(x^k - \gamma w^{k+1}) \end{cases}, \tag{4.1}$$

where $f^*$ is the conjugate function of $f$. Its resolvent can be evaluated by the Moreau's identity,

$$x = (\mathrm{I} + \gamma\partial f)^{-1}(x) + \gamma\left(\mathrm{I} + \frac{1}{\gamma}\partial f^*\right)^{-1}\left(\frac{x}{\gamma}\right).$$

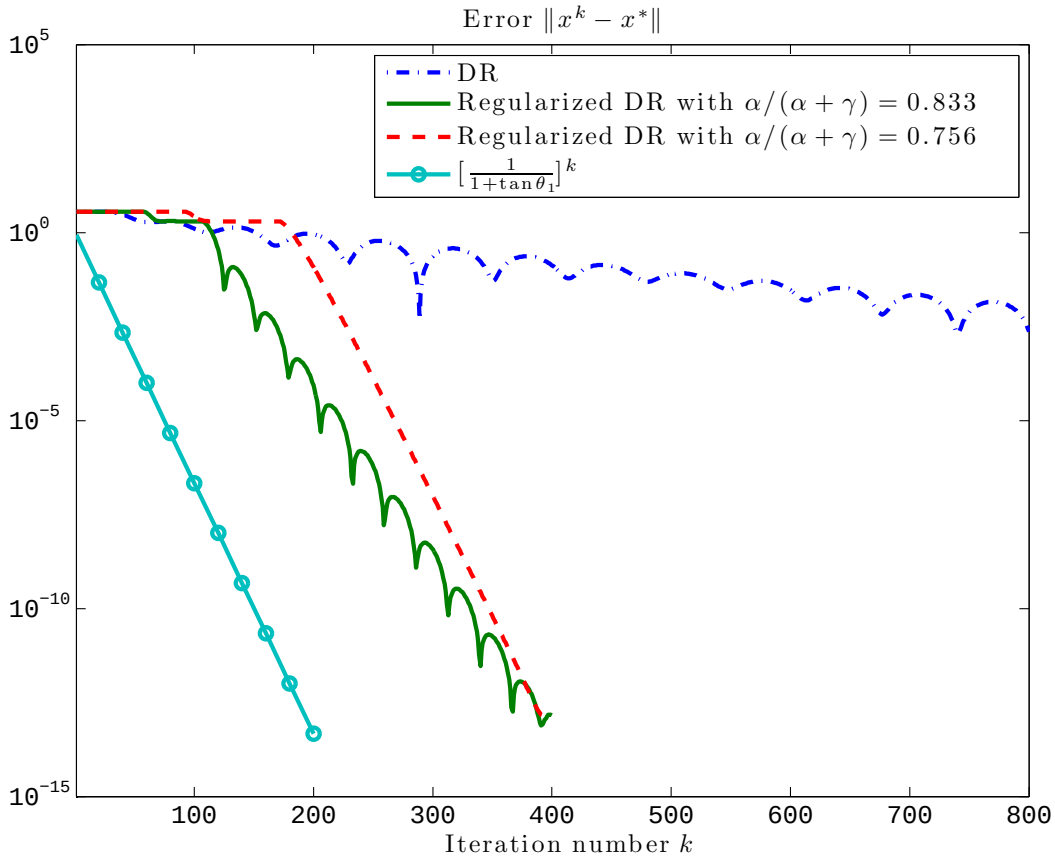



Figure 3.2: Example 5: $\alpha = 30$ is fixed. DR stands for (1.3) and Regularized DR stands for (3.2).

Let $X^n = \frac{1}{n}\sum_{k=1}^{n} x^k$ and $W^n = \frac{1}{n}\sum_{k=1}^{n} w^k$, then the duality gap of the point $(X^n, W^n)$ converges with the rate $\mathcal{O}(\frac{1}{n})$. See [6] for the proof. If $f(x) = \|x\|_1$ and $g(x) = \iota_{\{x:Ax=b\}}$, then $w^k$ will converge to a dual certificate $\eta \in \partial\|x^*\|_1 \cap \mathcal{R}(A^T)$.

## 4.2 Alternating direction method of multipliers

In this subsection we recall the the widely used *alternating direction method of multipliers* (ADMM), which serves as a preliminary for the next subsection. ADMM [14, 13] was shown in [12] to be equivalent to the Douglas-Rachford splitting on the dual problem. To be more specific, consider

$$\min_{z\in\mathbb{R}^m} \Psi(z) + \Phi(Dz), \tag{P}$$

where $\Psi$ and $\Phi$ are convex functions and $D$ is a $n \times m$ matrix. The dual problem of the equivalent constrained form $\min \Psi(z) + \Phi(w)$ s.t. $Dz = w$ is

$$\min_{x\in\mathbb{R}^n} \Psi^*(-D^T x) + \Phi^*(x). \tag{D}$$

By applying the Douglas-Rachford splitting (1.2) on $F = \partial[\Psi^* \circ (-D^T)]$ and $G = \partial\Phi^*$, one recovers the classical ADMM algorithm for (P),

$$\begin{cases} z^{k+1} = \arg\min_z \Psi(z) + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + Dz - w^k\|^2 \\ w^{k+1} = \arg\min_w \Phi(w) + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + Dz^{k+1} - w\|^2 \\ x^{k+1} = x^k + \gamma(Dz^{k+1} - w^{k+1}) \end{cases}, \tag{ADMM}$$

with the change of variable $y^k = x^k + \gamma w^k$, and $x^k$ unchanged.

After its discovery, ADMM has been regarded as a special augmented Lagrangian method. It turns out that ADMM can also be interpreted in the context of Bregman iterations. The split Bregman method [15] for (P) is exactly the same as (ADMM), see [23]. Since we are interested in Douglas-Rachford splitting for the primal formulation of the $\ell^1$ minimization, the algorithms analyzed in the previous sections are equivalent to ADMM or split Bregman method applied to the dual formulation.

## 4.3 Split Bregman method on the dual problem

In this subsection we show that the analysis in Section 3 can also be applied to the split Bregman method on the dual formulation [24]. The dual problem of $\ell^2$ regularized basis pursuit (3.1) can be written as

$$\min_z -b^T z + \frac{\alpha}{2}\|A^T z - \mathbb{P}_{[-1,1]^n}(A^T z)\|^2, \tag{4.2}$$

where $z$ denotes the dual variable, see [25].

By switching the first two lines in (ADMM), we get a slightly different version of ADMM:

$$\begin{cases} w^{k+1} = \arg\min_w \Phi(w) + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + Dz^k - w\|^2 \\ z^{k+1} = \arg\min_z \Psi(z) + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + Dz - w^{k+1}\|^2 \\ x^{k+1} = x^k + \gamma(Dz^{k+1} - w^{k+1}) \end{cases}. \tag{ADMM2}$$

The well-known equivalence between (ADMM) and Douglas-Rachford splitting was first explained in [12]. See also [23, 10]. For completeness, we discuss the equivalence between (ADMM2) and Douglas-Rachford splitting.

**Theorem 4.1.** *The iterates in (ADMM2) are equivalent to the Douglas-Rachford splitting (1.2) on $F = \partial\Phi^*$ and $G = \partial[\Psi^* \circ (-D^T)]$ with $y^k = x^{k-1} - \gamma w^k$.*

*Proof.* For any convex function $h$, we have $\lambda \in \partial h(p) \Longleftrightarrow p \in \partial h^*(\lambda)$, which implies

$$\hat{p} = \arg\min_p h(p) + \frac{\gamma}{2}\|Dp - q\|^2 \Longrightarrow \gamma(D\hat{p} - q) = J_{\gamma\partial(h^*\circ(-D^T))}(-\gamma q). \tag{4.3}$$

Applying (4.3) to the first two lines of (ADMM2), we get

$$x^k - \gamma w^{k+1} = J_{\gamma F}(x^k + \gamma D z^k) - \gamma D z^k. \tag{4.4}$$

$$x^k + \gamma D z^{k+1} - \gamma w^{k+1} = J_{\gamma G}(x^k - \gamma w^{k+1}). \tag{4.5}$$

Assuming $y^k = x^{k-1} - \gamma w^k$, we need to show that the $(k+1)$-th iterate of (ADMM2) satisfies $y^{k+1} = J_{\gamma F} \circ (2J_{\gamma G} - I)y^k + (I - J_{\gamma G})y^k$ and $x^{k+1} = J_{\gamma G}(y^{k+1})$.

Notice that (4.5) implies

$$J_{\gamma G}(y^k) = J_{\gamma G}(x^{k-1} - \gamma w^k) = x^{k-1} + \gamma D z^k - \gamma w^k.$$

So we have

$$J_{\gamma G}(y^k) - y^k = x^{k-1} + \gamma D z^k - \gamma w^k - (x^{k-1} - \gamma w^k) = \gamma D z^k,$$

and

$$2J_{\gamma G}(y^k) - y^k = x^{k-1} + 2\gamma D z^k - \gamma w^k = x^{k-1} + \gamma D z^k - \gamma w^k + \gamma D z^k = x^k + \gamma D z^k.$$

Thus (4.4) becomes

$$y^{k+1} = J_{\gamma F} \circ (2J_{\gamma G} - I)y^k + (I - J_{\gamma G})y^k.$$

And (4.5) is precisely $x^{k+1} = J_{\gamma G}(y^{k+1})$. □

Applying (ADMM2) on (4.2) with $\Psi(z) = -b^T z$, $\Phi(z) = \frac{\alpha}{2}\|z - \mathbb{P}_{[-1,1]^n}(z)\|^2$ and $D = A^T$, we recover the LB-SB algorithm in [24],

$$\begin{cases} w^{k+1} = \arg\min_w \frac{\alpha}{2}\|w - \mathbb{P}_{[-1,1]^n}(w)\|^2 + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + A^T z^k - w\|^2 \\ z^{k+1} = \arg\min_z -b^T z + \frac{\gamma}{2}\|\frac{1}{\gamma}x^k + A^T z - w^{k+1}\|^2 \\ x^{k+1} = x^k + \gamma(A^T z^{k+1} - w^{k+1}) \end{cases}. \tag{LB-SB}$$

It is straightforward to check that $\Psi^* \circ (-A)(x) = \iota_{\{x:Ax=b\}}$ and $\Phi^*(x) = \|x\|_1 + \frac{1}{2\alpha}\|x\|^2$. By Theorem 4.1, (LB-SB) is exactly the same as (3.2). Therefore, all the results in Section 3 hold for (LB-SB). In particular, the dependence of the eventual linear convergence rate of (LB-SB) on the parameters is governed by (3.4) as illustrated in Figure 3.1.

**Remark 4.2.** *Let $z^*$ be the minimizer of (4.2) then $\alpha S_1(A^T z^*)$ is the solution to (3.1), see [25]. So $t^k = \alpha S_1(A^T z^k)$ can be used as the approximation to $x^*$, the solution to (3.1), as suggested in [24]. By Theorem 4.1, we can see that $x^k$ will converge to $x^*$ too. And it is easy to see that $x^k$ satisfies the constraint $Ax^k = b$ in (3.2). But $t^k$ does not necessarily lie in the affine set $\{x : Ax = b\}$. Thus $\{t^k\}$ and $\{x^k\}$ are two completely different sequences even though they both can be used in practice.*

## 4.4 Practical relevance

To implement the algorithm exactly as presented earlier, the availability of $A^+$ is necessary. Algorithms such as (1.3) and (3.2), the same as (LB-SB), are not suitable if $(AA^T)^{-1}$ is prohibitive to obtain. On the other hand, there are quite a few important problems for which $(AA^T)^{-1}$ is cheap to compute and store in memory. For instance, $AA^T$ may be relatively small and is a well-conditioned matrix in typical compressive sensing problems. Another example is when $A^T$ represents a tight frame transform, for which $AA^T$ is the identity matrix.

As for the efficiency of (LB-SB), see [24] for the comparison of (LB-SB) with other state-of-the-art algorithms.

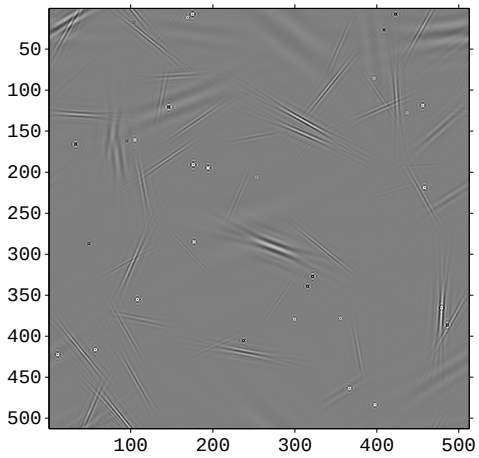
Next, we discuss several examples of (3.2) and (LB-SB) for the tight frame of discrete curvelets [4], in the scope of an application to interpolation of 2D seismic data. In the following examples, let $C$ denote the matrix representing the wrapping version of the two-dimensional fast discrete curvelet transform [4], then $C^T$ represents the inverse curvelet transform and $C^T C$ is the identity matrix since the curvelet transform is a tight frame.

**Example 6** We construct an example with $A = C^T$ to validate formula (3.4). Consider a random sparse vector $x^*$ with length 379831 and 93 nonzero entries, in the curvelet domain which is the range of the curvelet transform of $512 \times 512$ images. The size of the abstract matrix $C^T$ is $262144 \times 379831$. Notice that, for any $y \in \mathbb{R}^{512\times512}$, $Cy$ is implemented through fast Fourier transform, thus the explicit matrix representation of $C$ is never used in computation. Let $b = C^T x^*$ denote the $512 \times 512$ image generated by taking the inverse transform of $x^*$, see Figure 4.1 (a).

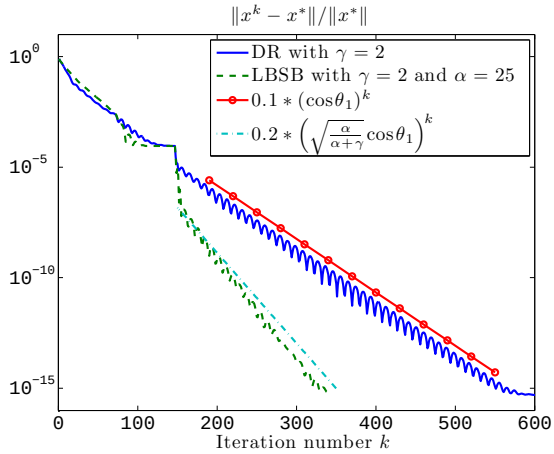
Suppose only the data $b$ is given, to recover a sparse curvelet coefficient, we can solve (1.1) with $A = C^T$ and $x$ being vectors in curvelet domain.

We use both (1.3) and (3.2) with $\gamma = 2$ and $\alpha = 25$ to solve (1.1). Since $A$ is a huge implicitly defined matrix, it is not straightforward to compute the angles exactly by SVD as in small matrices examples. Instead, we obtain approximately the first principal angle $\theta_1 = \arccos(0.9459)$ between $\mathcal{N}(A)$ and $\mathcal{N}(B)$ in a more efficient ad hoc way in Appendix C. Assuming $\cos\theta_1 = 0.9459$ and $\frac{\alpha}{\alpha+\gamma} = \frac{25}{27}$, if $y^k$ in (3.2) converged to a fixed point of the same type (interior or boundary fixed point) as $y^k$ in (1.3), the eventual linear rate of (3.2)

should be $\sqrt{\frac{\alpha}{\alpha+\gamma}}\cos\theta_1$ by (3.4). As we can see in Figure 4.1 (b), the error curve for (3.2) matched well with the eventual linear rate $\sqrt{\frac{\alpha}{\alpha+\gamma}}\cos\theta_1$.




(a) The data $b = C^T x^*$.




(b) Here $\cos\theta_1 = 0.9459$. DR stands for (1.3) and LBSB stands for (3.2) and (LB-SB).

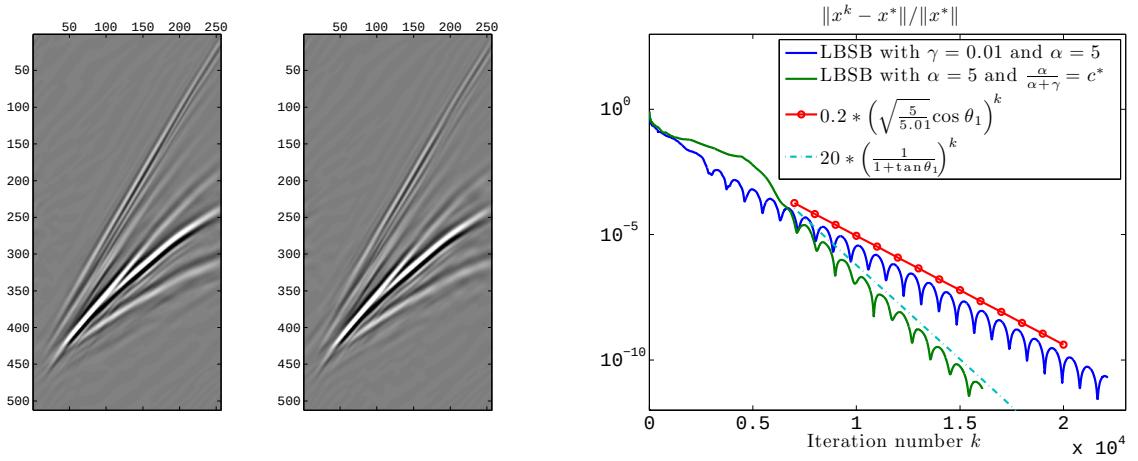
Figure 4.1: Example 6: recover a sparse curvelet coefficient.

**Example 7** In this example, we consider a more realistic data $b$ as shown in the left panel of Figure 4.2 (a). The data $b$ is generated by the following procedure. First, take a synthetic seismic dataset $\tilde{b}$ consisting of 256 traces (columns) and 512 time samples (rows). Second, solve the basis pursuit $\min_x \|x\|_1$ with $C^T x = \tilde{b}$ by (3.2) up to 50000 iterations. Third, set the entries in $x^{50000}$ smaller than $10^{-8}$ to zero and let $x^*$ denote the resulting sparse vector, which has 679 nonzero entries. Finally, set $b = C^T x^*$.

Given only the data $b$, the direct curvelet transform $Cb$ is not as sparse as $x^*$. Thus $Cb$ is not the most effective choice to compress the data. To recover the curvelet coefficient sequence $x^*$, we alternatively solve (1.1) with $A = C^T$ and $x$ being vectors in curvelet domain. For this particular example, $x^*$ is recovered. By the method in Appendix C, we get $\cos\theta_1 = 0.99985$. To achieve the best asymptotic rate, the parameter ratio $\frac{\alpha}{\alpha+\gamma}$ should be $c^* = \frac{1}{(\sin\theta_1+\cos\theta_1)^2} = 0.996549$ by (3.4). See Figure 4.2 (b) for the performance of (LB-SB) with fixed $\alpha = 5$ and we can see the asymptotic linear rate matches the best rate $\frac{1}{1+\tan\theta_1}$ when $\frac{\alpha}{\alpha+\gamma} = c^*$.

**Example 8** We consider an example of seismic data interpolation via curvelets. Let $b$ be the same data as in the previous example, see the left panel in Figure 4.2 (a). Let $\Omega$ be the sampling operator corresponding to 47 percent random traces missing, see Figure 4.3 (b).

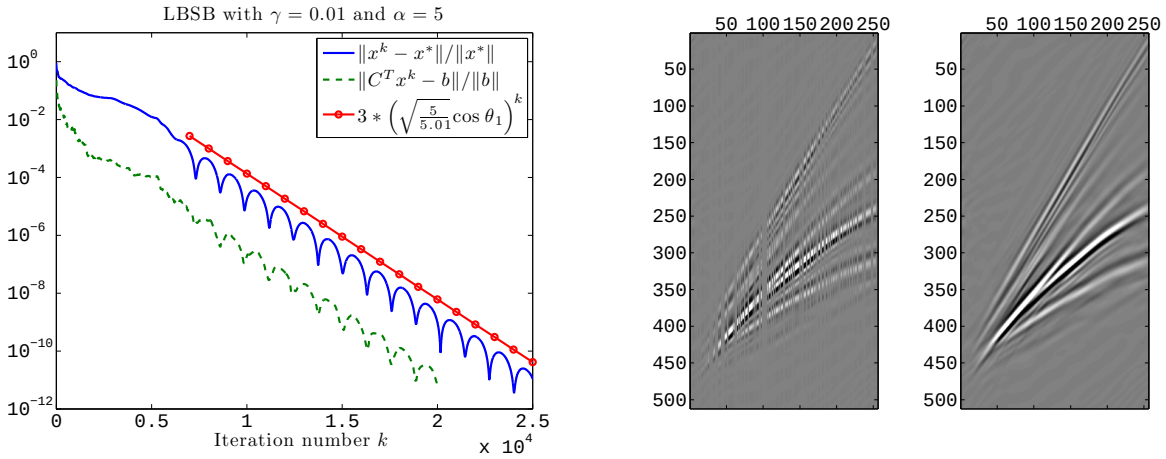
Given the observed data $\bar{b} = \Omega(b)$, to interpolate and recover missing data (traces), one effective model is to pursue sparsity in the curvelet domain [18], i.e., solving $\min_x \|x\|_1$ with the constraint $\Omega(C^T x) = \bar{b}$. Here $x$ is a vector of curvelet coefficients. If $x^*$ is a minimizer, then $C^T x^*$ can be used as the recovered data. Let $Ax = \Omega(C^T x)$. Then $A^+ = A^T$ since $\Omega$




(a) Left: the original data $b$. Right: reconstructed data with 300 largest curvelet coefficients $x^*$.

(b) The eventual linear convergence. LBSB stands for (3.2) and (LB-SB).

Figure 4.2: Example 7: compression of seismic data.




(a) The eventual linear convergence. LBSB stands for (3.2) and (LB-SB).

(b) Left: observed data, about 47% random traces missing. Right: recovered data after 200 iterations with relative error $\|C^T x^{200} - b\|/\|b\| = 2.6\%$ where $b$ is the original data in Figure 4.2 (a).

Figure 4.3: Example 8: seismic data interpolation.

represents a sampling operator. Thus (3.2) and (LB-SB) are straightforward to implement. For this relatively ideal example, the original data $b$ can be recovered. We also observe the eventual linear convergence. See Figure 4.3 (b) for the recovered data after 200 iterations of (3.2) and (LB-SB).

# 5 Conclusion

In this paper, we analyze the asymptotic convergence rate for Douglas-Rachford splitting algorithms on the primal formulation of the basis pursuit, providing a quantification of asymptotic convergence rate of such algorithms. In particular, we get a formula of the asymptotic convergence rate for $\ell^2$-regularized Douglas-Rachford, an algorithm equivalent to the dual split Bregman method. The explicit dependence of the convergence rate on the parameters may shed light on how to choose parameters in practice.

# Appendix A

**Lemma A.1.** *Let $T$ be a firmly non-expansive operator, i.e., $\|T(u)-T(v)\|^2 \leq \langle u-v, T(u)-T(v)\rangle$ for any $u$ and $v$. Then the iterates $y^{k+1}=T(y^k)$ satisfy $\|y^k-y^{k+1}\|^2 \leq \frac{1}{k+1}\|y^0-y^*\|^2$ where $y^*$ is any fixed point of $T$.*

*Proof.* The firm non-expansiveness implies

$$\begin{aligned}\|(I-T)(u)-(I-T)(v)\|^2 &= \|u-v\|^2+\|T(u)-T(v)\|^2-2\langle u-v, T(u)-T(v)\rangle\\ &\leq \|u-v\|^2-\|T(u)-T(v)\|^2.\end{aligned}$$

Let $u=y^*$ and $v=y^k$, then

$$\|y^{k+1}-y^k\|^2 \leq \|y^k-y^*\|^2-\|y^{k+1}-y^*\|^2.$$

Summing the inequality above, we get $\sum_{k=0}^{\infty}\|y^{k+1}-y^k\|^2 = \|y^0-y^*\|^2$. By the firm non-expansiveness and the Cauchy-Schwarz inequality, we have $\|y^{k+1}-y^k\| \leq \|y^k-y^{k-1}\|$, which implies $\|y^{n+1}-y^n\|^2 \leq \frac{1}{n+1}\sum_{k=0}^{n}\|y^{k+1}-y^k\|^2 \leq \frac{1}{n+1}\sum_{k=0}^{\infty}\|y^{k+1}-y^k\|^2 = \frac{1}{n+1}\|y^0-y^*\|^2$. □

For the Douglas-Rachford splitting, see [17] for a different proof for this fact.

# Appendix B

To apply Douglas-Rachford splitting (1.2) to basis pursuit (1.1), we can also choose $g(x)=\|x\|_1$ and $f(x)=\iota_{\{x:Ax=b\}}$, then Douglas-Rachford iterations become

$$\begin{cases} y^{k+1}=x^k+A^+(b-A(2x^k-y^k)) \\ x^{k+1}=S_\gamma(y^{k+1}) \end{cases}. \tag{B.1}$$

The discussion in Section 2 can be applied to (B.1). In particular, the corresponding matrix in (2.5) is $\mathbf{T} = (I_n - A^+A)(I_n - A^+A) + A^+AB^+B$, thus all the asymptotic convergence rates remain valid. For all the numerical tests in this paper, we did not observe any significant difference in performance between (1.3) and (B.1).

For the $\ell^2$ regularized basis pursuit (3.1), let $g(x) = \|x\|_1 + \frac{1}{2\alpha}\|x\|^2$ and $f(x) = \iota_{\{x:Ax=b\}}$, then a different Douglas-Rachford algorithm reads

$$\begin{cases} y^{k+1} = x^k + A^+(b - A(2x^k - y^k)) \\ x^{k+1} = \frac{\alpha}{\alpha+\gamma} S_\gamma(y^{k+1}) \end{cases}. \tag{B.2}$$

Similarly, the discussion in Section 3 holds for (B.2). In fact, there are more choices such as $g(x) = \|x\|_1$ and $f(x) = \iota_{\{x:Ax=b\}} + \frac{1}{2\alpha}\|x\|^2$ when using Douglas-Rachford splitting (1.2). We observed no significant differences in numerical performance between (3.2) and these variants including (B.2).

# Appendix C

Suppose $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$, we discuss an ad hoc way to find an approximation of the first principal angle $\theta_1$ between $\mathcal{N}(A)$ and $\mathcal{N}(B)$. Define the projection operators $P_{\mathcal{N}(A)}(x) = (I - A^+A)x$ and $P_{\mathcal{N}(B)}(x) = (I - B^+B)x$. Consider finding a point in the intersections of two linear subspaces,

$$\text{find} \quad x \in \mathcal{N}(A) \cap \mathcal{N}(B), \tag{C.1}$$

by von Neumann's alternating projection algorithm,

$$x^{k+1} = P_{\mathcal{N}(A)} P_{\mathcal{N}(B)}(x^k), \tag{C.2}$$

or the Douglas-Rachford splitting,

$$y^{k+1} = \frac{1}{2}[(2P_{\mathcal{N}(A)} - I)(2P_{\mathcal{N}(B)} - I) + I](y^k), \quad x^{k+1} = P_{\mathcal{N}(B)}(y^{k+1}). \tag{C.3}$$

For the algorithm (C.2), we have the error estimate $\|x^k\| = \|(I - A^+A)(I - B^+B)^k x^0\| \leq (\cos\theta_1)^{2k}\|x^0\|$ by (2.7).

Assume $y^*$ and $x^*$ are the fixed points of the iteration (C.3). Let $\mathbf{T} = (I - A^+A)(I - B^+B) + I$. For the algorithm (C.3), by (2.9), we have

$$\|x^{k+1} - x^*\| \leq \|y^{k+1} - y^*\| = \|\mathbf{T}(y^k - y^*)\| = \|\mathbf{T}^k(y^0 - y^*)\| \leq (\cos\theta_1)^k\|y^0 - y^*\|.$$

Notice that $\mathbf{0}$ is the only solution to (C.1). By fitting lines to $\log(\|x^k\|)$ for large $k$ in (C.2) and (C.3), we get an approximation of $2\log\cos\theta_1$ and $\log\cos\theta_1$ respectively. In pratice, (C.2) is better since the rate is faster and $\|x^k\|$ is monotone in $k$. This could be an efficient ad hoc way to obtain $\theta_1$ when the matrix $A$ is implicitly defined as in the examples in Section 4.4.

# References


[1] Francisco J. Aragón Artacho and Jonathan M. Borwein. Global convergence of a non-convex Douglas-Rachford iteration. *Journal of Global Optimization*, pages 1–17, 2012.

[2] Heinz H. Bauschke, Patrick L. Combettes, and D. Russell Luke. Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization. *J. Opt. Soc. Am. A*, 19(7):1334–1345, Jul 2002.

[3] Åke Björck and Gene H. Golub. Numerical Methods for Computing Angles Between Linear Subspaces. *Mathematics of Computation*, 27(123), 1973.

[4] E. Candes, L. Demanet, D. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Modeling Simulation*, 5(3):861–899, 2006.

[5] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203 – 4215, dec. 2005.

[6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, May 2011.

[7] Scott Shaobing Chen, David L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[8] Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53:475–504, 2004.

[9] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[10] E. Esser. Applications of Lagrangian based alternating direction methods and connections to split Bregman. CAM Report 09-31, UCLA, 2009.

[11] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *Information Theory, IEEE Transactions on*, 50(6):1341 – 1344, june 2004.

[12] D. Gabay. Applications of the method of multipliers to variational inequalities. *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems edited by M. FORTIN and R. GLOWINSKI*, 1983.

[13] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, January 1976.

[14] R. Glowinski and A. Marroco. Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualité d'une classe de problemes de dirichlet non lineares. *Revue Française d'Automatique, Informatique et Recherche Opérationelle*, 9:41–76, 1975.

[15] Tom Goldstein and Stanley Osher. The split Bregman method for L1-regularized problems. *SIAM J. Img. Sci.*, 2(2):323–343, April 2009.

[16] B. He and X. Yuan. On the O(1/$n$) convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[17] Bingsheng He and Xiaoming Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *preprint*, 2012.

[18] Felix J. Herrmann and Gilles Hennenfent. Non-parametric seismic data recovery with curvelet frames. *Geophysical Journal International*, 173(1):233–248, 2008.

[19] Robert Hesse and D. Russell Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *preprint*.

[20] M.-J. Lai and W. Yin. Augmented l1 and nuclear-norm models with a globally linearly convergent algorithm. Technical report, Rice University CAAM, 2012.

[21] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM J. on Optimization*, 13(3):702–725, August 2002.

[22] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16:964–979, 1979.

[23] Simon Setzer. Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage. In *Proceedings of the Second International Conference on Scale Space and Variational Methods in Computer Vision*, SSVM '09, pages 464–476, Berlin, Heidelberg, 2009. Springer-Verlag.

[24] Yi Yang, Michael Moller, and Stanley Osher. A dual split Bregman method for fast $\ell^1$ minimization. *Mathematics of Computation, to appear*.

[25] Wotao Yin. Analysis and generalizations of the linearized Bregman method. *SIAM J. Img. Sci.*, 3(4):856–877, October 2010.

[26] Wotao Yin and Stanley Osher. Error forgetting of bregman iteration. *Journal of Scientific Computing*, 54(2-3):684–695, 2013.

[27] Hui Zhang, Wotao Yin, and Lizhi Cheng. Necessary and sufficient conditions of solution uniqueness in $\ell 1$ minimization. Technical report, Rice University CAAM, 2012.