Accelerated Projected Steepest Descent Method for Nonlinear Inverse Problems with Sparsity Constraints

Gerd Teschke[*]   Claudia Borries[*]

July 3, 2009

Abstract

This paper is concerned with the construction of an iterative algorithm to solve nonlinear inverse problems with an ℓ_1 constraint. One extensively studied method to obtain a solution of such an ℓ_1 penalized problem is iterative soft-thresholding. Regrettably, such iteration schemes are computationally very intensive. A subtle alternative to iterative soft-thresholding is the projected gradient method that was quite recently proposed by Daubechies et.al. in [3]. The authors have shown that the proposed scheme is indeed numerically much thriftier. However, its current applicability is limited to linear inverse problems. In this paper we provide an extension of this approach to nonlinear problems. Adequately adapting the conditions on the (variable) thresholding parameter to the nonlinear nature, we can prove convergence in norm for this projected gradient method, with and without acceleration. A numerical verification is given in the context of nonlinear and non-ideal sensing. For this particular recovery problem we can achieve an impressive numerical performance (when comparing it to non-accelerated procedures).

Keywords: Nonlinear inverse problems, sparse recovery, steepest descent method, nonlinear and non-ideal sensing

Mathematics Subject Classification (2000): 15A29, 49M30, 65F22

1 Introduction

The main goal in this paper is the computation of an approximation to a solution of a nonlinear operator equation

$$F(x) = y \tag{1.1}$$

with an ℓ_1 constraint. Here we assume that $F : X \to Y$ is an possibly ill-posed operator between Hilbert spaces X and Y. In case only noisy data y^δ with $\|y^\delta - y\| \leq \delta$ are available, problem (1.1) has to be stabilized by regularization methods. In recent years, several iterative methods

to approximate/regularize the solution for linear ill-posed problems with sparsity constraints were developed, e.g. in [1, 2, 4, 5, 7, 8, 12] and a several more, and also extended to nonlinear problems, see, e.g., [10, 11]. The majority of these schemes are due to its simple nature very easy to use and can be applied in various reformulations to a very broad field of applications. However, in many practical situations one observes that Landweber/Richardson type iterations are rather slow and therefore are not competitive at all. This is especially for nonlinear problems the case, in which often (due to the implicitly given iterates) additional fixed point iterations are required.

The *classical Landweber iteration* for linear inverse problems (F linear and no sparsity constraint) is given by

$$x^{n+1} = x^n + \gamma F^*(y - Fx^n)$$

and can be viewed as a gradient descent algorithm with a fixed step size. This method is known to converge usually quite slowly; even the application of the nonlinear shrinkage operation,

$$x^{n+1} = S_\alpha(x^n + \gamma F^*(y - Fx^n)), \tag{1.2}$$

which comes into play by involving sparsity constraints, see [2], does not change this slow convergence. The same observation can be made in the nonlinear situation (F nonlinear) in which one possible variant of Landweber's iteration is given by

$$x^{n+1} = S_\alpha(x^n + \gamma F'(x^{n+1})^*(y - F(x^n))) , \tag{1.3}$$

for more details see [11]. One first serious step to accelerate these kind of iterations was suggested in [3]. In [3] the authors "borrowed a leaf" from standard linear steepest descent methods by using an adaptive step length. In addition, a detailed analysis of the characteristic dynamics of the classical thresholded Landweber iteration (1.2) has shown that the algorithm converges initially relatively fast, then it overshoots the ℓ_1 penalty, and it takes very long to re-correct back. A first intuitive way to circumvent this "external" detour is to force the iterates to remain within a particular ℓ_1 ball $B_R := \{x \in \ell_2; \|x\|_1 \leq R\}$. The authors of [3] proposed to achieve this by substituting the thresholding operation S_α by a projection P_{B_R}, where, for any closed convex set $\mathcal{C}$ and any x, the projection $P_{\mathcal{C}}(x)$ is defined as the unique point in $\mathcal{C}$ for which the ℓ_2 distance to x is minimal. With denoting P_{B_R} by P_R, this leads to the *projected Landweber iteration*,

$$x^{n+1} = P_R(x^n + \gamma F^*(y - Fx^n)). \tag{1.4}$$

However, the speed of convergence remained very slow. Therefore, as mentioned above, the authors suggested to introduce an adaptive "descent parameter" $\gamma^n > 0$ in each iteration leading to

$$x^{n+1} = P_R(x^n + \gamma^n F^*(y - Fx^n)). \tag{1.5}$$

The authors referred to this modified algorithm as the *projected gradient iteration* or the *projected steepest descent method*. They have determined how large one can choose the successive γ^n and have shown weak as well as strong convergence of the method (with and without acceleration) and gave numerical examples that clearly show the larger steps and faster convergence (when compared to (1.4)). Of course, there exist also other approaches for sparse recovery. Closely related to our approach are the methods presented in [9] and [14]. However, the analysis in [14] is limited to finite dimensions and the strategy provided in [9] is only suited for

linear inverse problems. The principle there is to reformulate the minimization problem as a bounded constrained quadratic program, and then apply iterative project gradient iterations.

The main concern in this paper is to extend the projected steepest descent method (1.5) to nonlinear problems. Since there are parallels in the construction process of (1.2) and (1.3), there is still hope that under similar assumptions on the nonlinear map F (that must also hold for the iterates in (1.3) to assure norm convergence) the projected steepest descent method converges in norm. Following the construction principles in [11], one natural projected steepest descent method for nonlinear operators might be given by

$$x^{n+1} = \mathrm{P}_R(x^n + \gamma^n F'(x^{n+1})^*(y - F(x^n))) . \tag{1.6}$$

As we shall see later in the paper this ansatz is indeed suitable and allows to prove many useful properties. Restricting ourselves to operators F that are twice Fréchet differentiable and for which all of the following three requirements (strong and Lipschitz continuity) hold true,

$$x^n \xrightarrow{w} x^\star \Longrightarrow F(x^n) \to F(x^\star) \tag{1.7}$$

$$F'(x^n)^*y \to F'(x^\star)^*y \quad \text{for all } y \tag{1.8}$$

$$\|F'(x) - F'(x')\| \le L\|x - x'\| , \tag{1.9}$$

we are able to prove strong convergence of the iterates in (1.6). It may happen that F already meets these conditions (1.8)-(1.9) as an operator from $X \to Y$. If not, this can be achieved by assuming more regularity of x, i.e. changing the domain of F a little. To this end, we assume that there exists a function space X^s, and a compact embedding operator $i^s : X^s \to X$. Then we may consider $\tilde F = F \circ i^s : X^s \longrightarrow Y$. Lipschitz regularity is preserved. Moreover, if now $x^n \xrightarrow{w} x^\star$ in X^s, then $x^n \to x^\star$ in X and, moreover, $\tilde F'(x^n) \to \tilde F'(x^\star)$ in the operator norm. This argument applies to arbitrary nonlinear continuous and Fréchet differentiable operators $F : X \to Y$ with continuous Lipschitz derivative as long as a function space X^s with compact embedding i^s into X is available.

At a first glance the made assumptions on F might seem to be somewhat restrictive. But compared to usually made assumptions in nonlinear inverse problems they are indeed reasonable and are fulfilled by numerous applications. At the cost of more technicalities several statements might also hold true for continuously Fréchet differentiable operators. However, for all the elaborated convergence analysis in this paper we cannot abstain from assumptions (1.8)-(1.9).

Another issue that is of great importance when dealing with ill-posed and inverse problems is to verify regularizing properties of the proposed method. Elaborations on this topic, however, are not provided within this but are planned for one of the subsequent papers. Nevertheless, we wish to briefly mention the theory that is still provided in the literature, which is so far unfortunately limited to linear problems, see, e.g., [6, Section 5.4]. Therefore, the concepts summarized in [6] not directly apply here and need to be extended. In any case, the question arises whether the convex constraint stabilize the problem or if it is still necessary to regularize the inverse problem. In general it seems to be meaningful to assume ill-posedness. Therefore, we need to introduce an additional stabilization. The iteration (1.6) can be viewed as iteration

3

scheme to approach the B_R-*best-approximate solution* $x_R^\dagger$, which we define as the minimizer of $\mathcal{D}(x) := \|F(x) - y\|^2$ on B_R, i.e.

$$\|F(x_R^\dagger) - y\| = \inf_x \{\|F(x) - y\|, \ x \in B_R\} \ \text{ and } \ \|x_R^\dagger\| = \min\{\|x\|, \ \|F(x) - y\| \wedge x \in B_R\} \ . \quad (1.10)$$

Since $x_R^\dagger \in B_R$, it is natural to require that the regularized solutions are in B_R as well. If $x^\dagger$ denotes the generalized solution of the unconstrained problem and if $x_R^\dagger = x^\dagger$, then all "standard results" concerning stability, convergence, and convergence rates hold also for the constrained case. If $x_R^\dagger \neq x^\dagger$, one might select a different regularization method, e.g.,

$$\min_{x \in B_R} \|F(x) - y\|^2 + \alpha \|x\|^2 \ .$$

For linear operators F regularization results are provided in [6]. Our situation would require an extension of the current established theory to nonlinear problems, but this is out-of-focus in this paper.

Organization of the remaining paper: In Section 2 we introduce the some standard notation and repeat some facts on convex analysis. In Section 3 we derive the necessary condition for critical points and we specify the conditions on the variable threshold parameter that allow us to prove strong convergence of iteration (1.6). Within the last section we discuss in greater detail a nonlinear and non-ideal sensing problem. Applying the proposed descent method, we can verify (numerically) perfect sparse recovery with an impressive numerical rate of convergence.

2 Preliminaries

We briefly introduce the some standard notation and repeat some facts that will be used in this paper. Let Λ be a countable index. We denote by $\ell_p(\Lambda)$ the space of sequences x for which $\|x\|_p < \infty$ (the usual sequence space norm). Assume we are given some Hilbert space X and some preassigned frame $\{\phi_\lambda\}_{\lambda \in \Lambda} \subset X$, i.e. there exist constants $0 < A \leq B < \infty$ such that for all $f \in X$,

$$A\|f\|_X^2 \leq \sum_{\lambda \in \Lambda} |\langle f, \phi_\lambda \rangle|^2 \leq B\|f\|_X^2 \ . \quad (2.1)$$

For this frame we may consider the so-called frame operator $S : X \to \ell_2(\Lambda)$ that is defined by $Sf = \{\langle f, \phi_\lambda \rangle\}_{\lambda \in \Lambda}$. Its adjoint S^* is given by $S^*x = \sum_{\lambda \in \Lambda} x_\lambda \phi_\lambda$. The functions ϕ_λ are typically linearly dependant but allow for a stable series expansion for any $f \in X$ of the form $f = S^*x$ for some $x \in \ell_2(\Lambda)$. The stability follows from (2.1). Due to $A \leq S^*S \leq B$, any function $f \in X$ can be reconstructed from its moments. Since a frame is overcomplete, there may exist many possibilities to represent f.

In what follows we focus on those problems in which the solution f has a sparse series expansion with respect to $\{\phi_\lambda\}_{\lambda \in \Lambda}$. This means that f can be written by a series expansion with only a very small number of non-vanishing coefficients x_λ, or that f is compressible (meaning that f can be nicely approximated by a sparse series expansion).

In (1.1) we have introduced the operator F as a map between Hilbert spaces X and Y. Typically X represents the Hilbert space of functions in which we are searching for a solution of

our inverse problem. But when it comes to numerical schemes to solve the inverse problem, we have to find a suitable expansion/representation of the solution. Since we aim here to apply the concept of frames, the goal of finding the function translates then into finding a corresponding sequence of coefficients x that is used in the associated series expansion to represent f. Therefore, it makes sense to consider F as a map between $\ell_2(\Lambda)$ and Y.

Before analyzing the proposed projected steepest descent (1.6), we provide some analysis of ℓ_2 projections onto ℓ_1 balls. The listed properties are proved here for completeness. They can be retraced in [3], from where they are partially taken, or to some extent in [4, 5].

Lemma 1 $\forall a \in \ell_2(\Lambda), \forall \tau > 0 : \|S_\tau(a)\|_1$ *is a piecewise linear, continuous, decreasing function of τ; moreover, if $a \in \ell_1(\Lambda)$ then $\|S_0(a)\|_1 = \|a\|_1$ and $\|S_\tau(a)\|_1 = 0$ for $\tau \geq \max_i |a_i|$.*

Proof. $\|S_\tau(a)\|_1 = \sum_{\lambda \in \Lambda} |S_\tau(a_\lambda)| = \sum_{\lambda \in \Lambda} S_\tau(|a_\lambda|) = \sum_{|a_\lambda| > \tau} (|a_\lambda| - \tau)$; the sum in the right hand side is finite for $\tau > 0$. $\square$

Lemma 2 *If $\|a\|_1 > R$, then the ℓ_2 projection of a on the ℓ_1 ball with radius R is given by $P_R(a) = S_\mu(a)$, where μ (depending on a and R) is chosen such that $\|S_\mu(a)\|_1 = R$. If $\|a\|_1 \leq R$ then $P_R(a) = S_0(a) = a$.*

Proof. Suppose $\|a\|_1 > R$. By Lemma 1: $\exists \mu > 0 : \|S_\mu(a)\|_1 = R$. Furthermore, $b = S_\mu(a)$ is the unique minimizer of $\|x - a\|^2 + 2\mu\|x\|_1$, i.e.,

$$\forall x \neq b : \qquad \|b - a\|^2 + 2\mu\|b\|_1 < \|x - a\|^2 + 2\mu\|x\|_1$$

Since $\|b\|_1 = R$, it follows that

$$\forall x \in B_R, x \neq b : \qquad \|b - a\|^2 < \|x - a\|^2$$

Hence b is closer to a than any other x in B_R. In other words, $P_R(a) = b = S_\mu(a)$. $\square$

Finally, P_R has the following additional properties:

Lemma 3 $\forall x \in \ell_2(\Lambda)$, $P_R(x)$ *is characterized as the unique vector in B_R such that*

$$\langle w - P_R(x), x - P_R(x) \rangle \leq 0, \qquad \forall w \in B_R \tag{2.2}$$

Moreover, the projection P_R is non-expansive:

$$\|P_R(x) - P_R(x')\| \leq \|x - x'\| \qquad \forall x, x' \in \ell_2(\Lambda) \tag{2.3}$$

Proof. Because B_R is convex, $(1 - t)P_R(x) + tw \in B_R$, $\forall w \in B_R$ and $t \in [0, 1]$. It follows that $\|x - P_R(x)\|^2 \leq \|x - [(1 - t)P_R(x) + tw]\|^2$; $\forall t \in [0, 1]$. This implies

$$\forall t \in [0, 1] : \qquad 0 \leq -2t\langle w - P_R(x), x - P_R(x) \rangle + t^2\|w - P_R(x)\|^2$$

5

It follows that

$$\langle w - \mathrm{P}_R(x), x - \mathrm{P}_R(x) \rangle \leq 0$$

which proves 2.2. Setting $w = \mathrm{P}_R(x')$ in 2.2, we get, for all x, x',

$$\langle \mathrm{P}_R(x') - \mathrm{P}_R(x), x - \mathrm{P}_R(x) \rangle \leq 0$$

Switching the role of x and x' one finds:

$$\langle \mathrm{P}_R(x') - \mathrm{P}_R(x), x' - \mathrm{P}_R(x') \rangle \geq 0$$

By combining these last two inequalities, one finds:

$$\langle \mathrm{P}_R(x') - \mathrm{P}_R(x), x' - x - \mathrm{P}_R(x') + \mathrm{P}_R(x) \rangle \geq 0$$

or

$$\|\mathrm{P}_R(x') - \mathrm{P}_R(x)\|^2 \leq \langle \mathrm{P}_R(x') - \mathrm{P}_R(x), x' - x \rangle$$

by Cauchy-Schwarz this gives

$$\|\mathrm{P}_R(x') - \mathrm{P}_R(x)\|^2 \leq \langle \mathrm{P}_R(x') - \mathrm{P}_R(x), x' - x \rangle \leq \|\mathrm{P}_R(x') - \mathrm{P}_R(x)\| \|x' - x\|$$

from which inequality 2.3 follows. $\qquad\square$

Lemma 1 and 2 provide a simple recipe for computing the projection $\mathrm{P}_R(a)$. First, sort the absolute values of the components of a (an $\mathcal{O}(m \log m)$ operation if $\#\Lambda = m$ is finite), resulting in the rearranged sequence $(a_l^*)_{l=1,\dots,m}$, with $a_l^* \geq a_{l+1}^* \geq 0; \forall l$. Next, perform a search to find k such that

$$\|\mathrm{S}_{a_k^*}(a)\|_1 = \sum_{l=1}^{k-1}(a_l^* - a_k^*) \leq R < \sum_{l=1}^{k}(a_l^* - a_{k+1}^*) = \|\mathrm{S}_{a_{k+1}^*}(a)\|_1 \ .$$

The complexity of this step is again $\mathcal{O}(m \log m)$. Finally, set $\nu := k^{-1}(R - \|S_{a_k^*}(a)\|_1)$, and $\mu := a_k^* - \nu$. Then

$$
\begin{aligned}
\|\mathrm{S}_\mu(a)\|_1 &= \sum_{i \in \Lambda} \max(|a_i| - \mu, 0) = \sum_{l=1}^{k}(a_l^* - \mu) \\
&= \sum_{l=1}^{k-1}(a_l^* - a_k^*) + k\nu = \|\mathrm{S}_{a_k^*}(a)\|_1 + k\nu = R \ .
\end{aligned}
$$

3 Projected Steepest Descent and Convergence

We have now collected some facts on the projector P_R and on convex analysis issues that allow for convergence analysis of the projected steepest descent method defined in (1.5). In what follows, we essentially proceed as in [3]. But as we shall see, several serious technical changes (including also a weakening of a few statements) but also significant extensions of the nice analysis provided in [3] need to be made. For instance, due to the nonlinearity of F, several uniqueness statements proved in [3] carry not over in its full glory. Nevertheless, the main propositions on *weak* and *strong convergence* can be achieved (of course, at the cost of involving much more technicalities).

3.1 Necessary Condition

Lemma 4 *If the vector $\tilde{x}_R \in \ell_2(\Lambda)$ is a minimizer of $\mathcal{D}(x) = \|F(x) - y\|^2$ on B_R then for any $\gamma > 0$,*

$$\mathrm{P}_R(\tilde{x}_R + \gamma F'(\tilde{x}_R)^*(y - F(\tilde{x}_R))) = \tilde{x}_R \ ,$$

which is equivalent to

$$\langle F'(\tilde{x}_R)^*(y - F(\tilde{x}_R)), w - \tilde{x}_R \rangle \leq 0, \quad for\ all\ w \in B_R \ .$$

Proof. Since F is twice Fréchet differentiable, we have the following Taylor expansion

$$F(x + h) = F(x) + F'(x)h + R(x, h) \quad \text{with} \ \ \|R(x, h)\| \leq \frac{L}{2}\|h\|^2 \ .$$

If now $\tilde{x}_R$ minimizes $\mathcal{D}$ on B_R, then for all $w \in B_R$ and all $t \in [0, 1]$,

$$
\begin{aligned}
D(\tilde{x}_R) &\leq \mathcal{D}((1 - t)\tilde{x}_R + tw) = \mathcal{D}(\tilde{x}_R + t(w - \tilde{x}_R)) \\
&= \|F(\tilde{x}_R + t(w - \tilde{x}_R)) - y\|^2 = \|F(\tilde{x}_R) - y + F'(\tilde{x}_R)t(w - \tilde{x}_R) + R(\tilde{x}_R, t(w - \tilde{x}_R))\|^2 \\
&= \mathcal{D}(\tilde{x}_R) + 2\langle F'(\tilde{x}_R)^*(F(\tilde{x}_R) - y), t(w - \tilde{x}_R) \rangle + 2\langle F(\tilde{x}_R) - y, R(\tilde{x}_R, t(w - \tilde{x}_R)) \rangle \\
&\quad + \|F'(\tilde{x}_R)t(w - \tilde{x}_R) + R(\tilde{x}_R, t(w - \tilde{x}_R))\|^2 \\
&\leq \mathcal{D}(\tilde{x}_R) + 2t\langle F'(\tilde{x}_R)^*(F(\tilde{x}_R) - y), w - \tilde{x}_R \rangle \\
&\quad + t^2\|w - \tilde{x}_R\|^2(L\|F(\tilde{x}_R) - y\| + 2\|F'(\tilde{x}_R)\|) + t^4 L\|w - \tilde{x}_R\|^4 \ .
\end{aligned}
$$

This implies

$$\langle F'(\tilde{x}_R)^*(y - F(\tilde{x}_R)), w - \tilde{x}_R \rangle \leq 0 \ ,$$

and therefore, for all $\gamma > 0$,

$$\langle \tilde{x}_R + \gamma F'(\tilde{x}_R)^*(y - F(\tilde{x}_R)) - \tilde{x}_R, w - \tilde{x}_R \rangle \leq 0 \ .$$

By Lemma 3 this implies the assertion. $\qquad\square$

Lemma 4 provides just a necessary condition for a minimizer $\tilde{x}_R$ of $\mathcal{D}$ on B_R. The minimizers of $\mathcal{D}$ on B_R need not be unique. Nevertheless, we have

Lemma 5 *If $\tilde{x}, \tilde{\tilde{x}} \in B_R$, if $\tilde{x}$ minimizes $\mathcal{D}$ and if $\tilde{x} - \tilde{\tilde{x}} \in \ker F'(w)$ for all $w \in B_R$ then $\tilde{\tilde{x}}$ minimizes $\mathcal{D}$ as well.*

Proof. This statement is easy to achieve,

$$
\begin{aligned}
D(\tilde{x}) &= \|F(\tilde{x} - \tilde{\tilde{x}} + \tilde{\tilde{x}}) - y\|^2 \\
&= \|F(\tilde{\tilde{x}}) - y + \int_0^1 F'(\tilde{\tilde{x}} + \tau(\tilde{x} - \tilde{\tilde{x}}))(\tilde{x} - \tilde{\tilde{x}})d\tau\|^2 = D(\tilde{\tilde{x}}) \ .
\end{aligned}
$$

$\qquad\square$

3.2 Weak Convergence

In order to achieve convergence results, we have to specify the choice of γ^n. To this end, we first introduce the constant r,

$$r := \max\{2 \sup_{x \in B_R} \|F'(x)\|^2 \,,\, 2L\sqrt{\mathcal{D}(x^0)}\}\,, \tag{3.1}$$

where x_0 denotes a first initial guess for the solution to be reconstructed. One role of the constant r can be seen in the following estimate which is possible by the first order Taylor expansion of F,

$$\|F(x^{n+1}) - F(x^n)\|^2 \leq \sup_{x \in B_R} \|F'(x)\|^2 \|x^{n+1} - x^n\|^2 \leq \frac{r}{2}\|x^{n+1} - x^n\|^2\,.$$

With the help of (3.1) we define a sequence of real numbers which we denote by β^n.

Definition 1 *We say that the sequence $\{\beta^n\}_{n \in \mathbb{N}}$ satisfies Condition (B) with respect to the sequence $\{x^n\}_{n \in \mathbb{N}}$ if there exists n_0 such that:*

$$(B1) \qquad \bar{\beta} := \sup\{\beta^n; n \in \mathbb{N}\} < \infty \quad \text{and} \quad \inf\{\beta^n; n \in \mathbb{N}\} \geq 1$$

$$(B2) \qquad \beta^n\|F(x^{n+1}) - F(x^n)\|^2 \leq \frac{r}{2}\|x^{n+1} - x^n\|^2 \quad \forall n \geq n_0$$

$$(B3) \qquad \beta^n L\sqrt{\mathcal{D}(x^n)} \leq \frac{r}{2}\,.$$

By condition (B1) we ensure

$$\|F(x^{n+1}) - F(x^n)\|^2 \leq \beta^n \|F(x^{n+1} - F(x^n)\|^2\,.$$

The idea of adding condition (B2) is to find the largest number $\beta^n \geq 1$ such that

$$0 \leq -\|F(x^{n+1}) - F(x^n)\|^2 + \frac{r}{2\beta^n}\|x^{n+1} - x^n\|^2$$

is as small as possible. The reason can be verified below in the definition of the gaussian surrogate functional Φ_β in Lemma 6. The goal is to ensure that Φ_{β^n} is not too far off $\mathcal{D}(x^n)$. The additional restriction (B3) was introduced to ensure convexity of Φ_{β^n} and convergence of the fixed point map Ψ in Lemma 7 (as we will prove below).

Because the definition of x^{n+1} involves β^n and vice versa, the inequality (B2) has an implicit quality. In practice, it is not straightforward to pick β^n adequately. This issue will be discussed later in Subsection 3.4.

In the remaining part of this subsection we prove weak convergence of any subsequence of $\{x^n\}_{n \in \mathbb{N}}$ towards weak limits that fulfill the necessary condition for minimizers of $\mathcal{D}$ on B_R.

Lemma 6 *Assume F to be twice Fréchet differentiable and $\beta \geq 1$. For arbitrary fixed $x \in B_R$ assume $\beta L\sqrt{\mathcal{D}(x)} \leq r/2$ and define the functional $\Phi_\beta(\cdot, x)$ by*

$$\Phi_\beta(w, x) := \|F(w) - y\|^2 - \|F(w) - F(x)\|^2 + \frac{r}{\beta}\|w - x\|^2\,. \tag{3.2}$$

Then there exists a unique $w \in B_R$ that minimizes the restriction to B_R of $\Phi_\beta(w, x)$. We denote this minimizer by $\hat{w}$ which is given by

$$\hat{w} = \mathrm{P}_R\left(x + \frac{\beta}{r}F'(\hat{w})^*(y - F(x))\right).$$

Proof. First, we prove that if F is twice Fréchet differentiable then $\Phi_\beta(\cdot, x)$ is strictly convex. To simplify the notation, we define

$$J(w) := \Phi_\beta(w, x) = \|F(w) - y\|^2 - \|F(w) - F(x)\|^2 + \frac{r}{\beta}\|w - x\|^2.$$

For strict convexity, we have to show for all $w, w' \in B_R$ and all $t \in (0, 1)$ that

$$J((1 - t)w + tw') < (1 - t)J(w) + tJ(w').$$

With the help of the second order Taylor expansion for F, we observe

$$
\begin{aligned}
J(w + h) &= \|F(w + h) - y\|^2 - \|F(w + h) - F(x)\|^2 + \frac{r}{\beta}\|w + h - x\|^2 \\
&= \|F(w) - y + F'(w)h + R(w, h)\|^2 - \|F(w) - F(x) + F'(w)h + R(w, h)\|^2 \\
&\quad + \frac{r}{\beta}\|w + h - x\|^2 \\
&= J(w) + 2\{\langle F(w) - y, F'(w)h\rangle - \langle F(w) - F(x), F'(w)h\rangle + \frac{r}{\beta}\langle w - x, h\rangle\} \\
&\quad + 2\langle F(x) - y, R(w, h)\rangle + \frac{r}{\beta}\|h\|^2 \\
&= J(w) + J'(w)h + \rho(w, h),
\end{aligned}
$$

where

$$\rho(w, h) := 2\langle F(x) - y, R(w, h)\rangle + \frac{r}{\beta}\|h\|^2. \tag{3.3}$$

Therefore we have

$$
\begin{aligned}
J((1 - t)w + tw') &= J(w + t(w' - w)) = J(w' + (1 - t)(w - w')) \\
&= (1 - t)J(w + t(w' - w)) + tJ(w' + (1 - t)(w - w')) \\
&= (1 - t)\left(J(w) + tJ'(w)(w' - w) + \rho(w, t(w' - w))\right) \\
&\quad + t\left(J(w') + (1 - t)J'(w')(w - w') + \rho(w', (1 - t)(w - w'))\right) \\
&= (1 - t)J(w) + tJ(w') + D(w, w', t).
\end{aligned}
$$

The functional J is now strictly convex, if for all $w, w' \in B_R$ and all $t \in (0, 1)$,

$$
\begin{aligned}
D(w, w', t) &:= t(1 - t)(J'(w) - J'(w'))(w' - w) \\
&\quad + (1 - t)\rho(w, t(w' - w)) + t\rho(w', (1 - t)(w - w')) < 0.
\end{aligned}
$$

We have

$$(J'(w) - J'(w'))(w' - w) = -\frac{2r}{\beta}\|w - w'\|^2 - 2\langle y - F(x), (F'(w) - F'(w'))(w' - w)\rangle.$$

Moreover, as F is twice differentiable,

$$F'(w) = F'(w') + \int_0^1 F''(w' + \tau(w - w'))(w - w', \cdot)d\tau$$

and consequently,

$$(J'(w) - J'(w'))(w' - w) = -\frac{2r}{\beta}\|w - w'\|^2 + 2\langle y - F(x), \int_0^1 F''(w' + \tau(w - w'))(w - w')^2 d\tau\rangle . \quad (3.4)$$

Again, by Fréchet differentiability, $R(x, h)$ in (3.3) is given by

$$R(x, h) = \int_0^1 (1 - \tau)F''(x + \tau h)(h)^2 \, d\tau \ ,$$

and consequently we obtain

$$\begin{aligned}
R(w, t(w' - w)) &= t^2 \int_0^1 (1 - \tau)F''(w + \tau t(w' - w))(w' - w)^2 \, d\tau \\
&= \int_{1-t}^1 (\tau - (1 - t))F''(w' + \tau(w - w'))(w - w')^2 \, d\tau
\end{aligned}$$

$$(3.5)$$

and by same arguments

$$R(w', (1 - t)(w - w')) = \int_0^{1-t} (1 - t - \tau)F''(w' + \tau(w - w'))(w - w')^2 \, d\tau . \quad (3.6)$$

Combining definition (3.3) and equations (3.4), (3.5) and (3.6) yields

$$D(w, w', t) = -t(1 - t)\frac{r}{\beta}\|w - w'\|^2 + 2\langle y - F(x), I(w, w', t)\rangle , \quad (3.7)$$

where

$$\begin{aligned}
I(w, w', t) \ := \ & t(1 - t)\int_0^1 F''(w' + \tau(w - w'))(w - w')^2 \, d\tau \\
& -(1 - t)\int_{1-t}^1 (\tau - (1 - t))F''(w' + \tau(w - w'))(w - w')^2 \, d\tau \\
& -t\int_0^{1-t} (1 - t - \tau)F''(w' + \tau(w - w'))(w - w')^2 \, d\tau .
\end{aligned}$$

10

The functional $I(w, w', t)$ can now be recast as follows

$$
I(w, w', t) \;=\; t \int_0^{1-t} \tau F''(w' + \tau(w - w'))(w - w')^2 \, d\tau
$$

$$
+ (1 - t) \int_{1-t}^{1} (1 - \tau) F''(w' + \tau(w - w'))(w - w')^2 \, d\tau .
$$

In order to estimate $\|I(w, w', t)\|$ it is necessary to estimate the integrals separately. Due to the Lipschitz–continuity of the first derivative, the second derivative can be globally estimated by L, and it follows,

$$
t \left\| \int_0^{1-t} \tau F''(w' + \tau(w - w'))(w - w')^2 \, d\tau \right\| \;\leq\; t \frac{(1-t)^2}{2} L \|w - w'\|^2 ,
$$

$$
(1 - t) \left\| \int_{1-t}^{1} (1 - \tau) F''(w' + \tau(w - w'))(w - w')^2 \, d\tau \right\| \;\leq\; (1 - t) \frac{t^2}{2} L \|w - w'\|^2
$$

and therefore

$$
\|I(w, w', t)\| \leq \frac{t(1 - t)}{2} L \|w - w'\|^2 . \tag{3.8}
$$

Combining (3.7), (3.8), and assumption $\beta L \sqrt{\mathcal{D}(x)} \leq r/2$, yields for $\lambda \in (0, 1)$

$$
\begin{aligned}
D(w, w', t) \;\leq\;& -t(1 - t) \frac{r}{\beta} \|w - w'\|^2 + 2\|y - F(x)\| \|I(w, w', t)\| \\
\leq\;& -t(1 - t) \frac{r}{\beta} \|w - w'\|^2 + \frac{t(1 - t)}{2} 2L \|y - F(x)\| \|w - w'\|^2 \\
\leq\;& -t(1 - t) \frac{r}{2\beta} \|w - w'\|^2 < 0 ,
\end{aligned}
$$

and thus the functional $\Phi_\beta(w, x)$ is strictly convex in w. Therefore there exists a unique minimizer $\hat{x}$ and thus we have for all $w \in B_R$ and all $t \in [0, 1]$

$$
\Phi_\beta(\hat{x}, x) \leq \Phi_\beta(\hat{x} + t(w - \hat{x}))
$$

and therefore

$$
\begin{aligned}
0 \;\leq\;& J(\hat{x} + t(w - \hat{x})) - J(\hat{x}) = t J'(\hat{x})(w - \hat{x}) + \rho(\hat{x}, t(w - \hat{x})) \\
=\;& 2t \langle F(x) - y, F'(\hat{x})(w - \hat{x}) \rangle + 2t \frac{r}{\beta} \langle \hat{x} - x, w - \hat{x} \rangle \\
& + 2 \langle F(x) - y, R(\hat{x}, t(w - \hat{x})) \rangle + \frac{r}{\beta} \|t(w - \hat{x})\|^2 \\
\leq\;& 2t \left\{ \langle F(x) - y, F'(\hat{x})(w - \hat{x}) \rangle + \frac{r}{\beta} \langle \hat{x} - x, w - \hat{x} \rangle \right\} \\
& t^2 \left\{ 2 \frac{r}{2\beta L} \frac{L}{2} \|w - \hat{x}\|^2 + \frac{r}{\beta} \|w - \hat{x}\|^2 \right\} .
\end{aligned}
$$

11

This implies for all $t \in [0, 1]$

$$0 \leq \left\{ \frac{\beta}{r} \langle F(x) - y, F'(\hat{x})(w - \hat{x}) \rangle + \langle \hat{x} - x, w - \hat{x} \rangle \right\} + \frac{3t}{4} \|w - \hat{x}\|^2 .$$

Consequently, we deduce

$$\langle x + \frac{\beta}{r} F'(\hat{x})^*(y - F(x)) - \hat{x}, w - \hat{x} \rangle \leq 0$$

which is by Lemma 3 equivalent to

$$\hat{x} = \mathrm{P}_R \left(x + \frac{\beta}{r} F'(\hat{x})^*(y - F(x)) \right)$$

and the proof is complete. □

The unique minimizer $\hat{x}$ is only implicitly given. We propose to apply a simple fixed point iteration to derive $\hat{x}$. The next lemma verifies that the corresponding fixed point map is indeed contractive and can therefore be used.

Lemma 7 *Assume* $\beta L \sqrt{\mathcal{D}(x)} \leq r/2$. *Then the map* $\Psi(\hat{x}) := \mathrm{P}_R(x + \beta/r F'(\hat{x})^*(y - F(x)))$ *is contractive and therefore the fixed point iteration*

$$\hat{x}^{l+1} = \Psi(\hat{x}^l)$$

converges to a unique fixed point.

Proof. Since P_R is non-expansive and F' is Lipschitz continuous, we have for all w, w',

$$\begin{aligned}
\|\Psi(w) - \Psi(w')\| &\leq \frac{\beta}{r} \|(F'(w)^* - F'(w')^*)(y - F(x))\| \leq \frac{\beta}{r} L \sqrt{\mathcal{D}(x)} \|w - w'\| \\
&\leq \frac{1}{2} \|w - w'\| < \|w - w'\| .
\end{aligned}$$

□

An immediate consequence of Lemma 6 is

Lemma 8 *Assume* x^{n+1} *is given by*

$$x^{n+1} = \mathrm{P}_R \left(x^n + \frac{\beta^n}{r} F'(x^{n+1})^*(y - F(x^n)) \right) ,$$

where r is as in (3.1) *and the β^n satisfy Condition* (B) *with respect to* $\{x^n\}_{n \in \mathbb{N}}$, *then the sequence* $\mathcal{D}(x^n)$ *is monotonically decreasing and* $\lim_{n \to \infty} \|x^{n+1} - x^n\| = 0$.

Proof. Comparing the definition of x^{n+1} and the statement in Lemma 6, we have that

$$x^{n+1} = \arg \min_x \Phi_{\beta^n}(x, x^n) .$$

We have

$$
\begin{aligned}
\mathcal{D}(x^{n+1}) &\leq D(x^{n+1}) + \|F(x^{n+1}) - F(x^n)\|^2 \\
&= \|F(x^{n+1}) - y\|^2 + 2\|F(x^{n+1}) - F(x^n)\|^2 - \|F(x^{n+1}) - F(x^n)\|^2 \\
&\leq \|F(x^{n+1}) - y\|^2 + \frac{r}{\beta^n}\|x^{n+1} - x^n\|^2 - \|F(x^{n+1}) - F(x^n)\|^2 \\
&= \Phi_{\beta^n}(x^{n+1}, x^n) \leq \Phi_{\beta^n}(x^n, x^n) = \mathcal{D}(x^n) \ .
\end{aligned}
$$

The second assertion can be shown with the help of (B2)

$$
\begin{aligned}
-\Phi_{\beta^{n+1}}(x^{n+1}, x^{n+1}) + \Phi_{\beta^n}(x^{n+1}, x^n) &= \frac{r}{\beta^n}\|x^{n+1} - x^n\|^2 - \|F(x^{n+1}) - F(x^n)\|^2 \\
&\geq \frac{r}{2\beta^n}\|x^{n+1} - x^n\|^2 \geq \frac{r}{2\bar{\beta}}\|x^{n+1} - x^n\|^2 \ .
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\sum_{n=0}^{N}\|x^{n+1} - x^n\|^2 &\leq \frac{2\bar{\beta}}{r}\sum_{n=0}^{N}\left\{\Phi_{\beta^n}(x^{n+1}, x^n) - \Phi_{\beta^{n+1}}(x^{n+1}, x^{n+1})\right\} \\
&\leq \frac{2\bar{\beta}}{r}\sum_{n=0}^{N}\left\{\Phi_{\beta^n}(x^n, x^n) - \Phi_{\beta^{n+1}}(x^{n+1}, x^{n+1})\right\} \\
&\leq \frac{2\bar{\beta}}{r}\left\{\Phi_{\beta^0}(x^0, x^0) - \Phi_{\beta^{N+1}}(x^{N+1}, x^{N+1})\right\} \leq \frac{2\bar{\beta}}{r}\Phi_{\beta^0}(x^0, x^0) \ .
\end{aligned}
$$

Since the bound does not depend on N, the infinite series $\sum_{n=0}^{\infty}\|x^{n+1} - x^n\|^2$ is bounded and thus $\lim_{n\to\infty}\|x^{n+1} - x^n\|^2 = 0$. $\qquad\square$

Since for all the iterates we have by definition $x^n \in B_R$, we automatically have $\|x^n\|_2 \leq R$ for all $n \in \mathbb{N}$. Therefore, the sequence $\{x^n\}_{n\in\mathbb{N}}$ must have weak accumulation points.

Proposition 9 *If $x^\star$ is a weak accumulation point of $\{x^n\}_{n\in\mathbb{N}}$, then it fulfills the necessary condition for a minimum of $\mathcal{D}(x)$ on B_R, i.e. for all $w \in B_R$,*

$$
\langle F'(x^\star)^*(y - F(x^\star)), w - x^\star\rangle \leq 0 \ .
$$

Proof. Since $x^{n_j} \xrightarrow{w} x^\star$, we have for fixed x and a

$$
\langle F'(x)x^{n_j}, a\rangle = \langle x^{n_j}, F'(x)^*a\rangle \longrightarrow \langle x^\star, F'(x)^*a\rangle = \langle F'(x)x^\star, a\rangle
$$

and therefore

$$
F'(x)x^{n_j} \xrightarrow{w} F'(x)x^\star. \tag{3.9}
$$

Due to Lemma 8, we also have $x^{n_j+1} \xrightarrow{w} x^\star$. Now we are prepared to show the necessary condition for the weak accumulation point $x^\star$. As the iteration is given by

$$
x^{n+1} = \mathrm{P}_R(x^n + \beta^n/rF'(x^{n+1})^*(y - F(x^n))) \ ,
$$

13

we have

$$\langle x^n + \beta^n / r F'(x^{n+1})^*(y - F(x^n)) - x^{n+1}, w - x^{n+1} \rangle \leq 0 \quad \text{for all } w \in B_R .$$

Specializing this inequality to the subsequence $\{x^{n_j}\}_{j \in \mathbb{N}}$ yields

$$\langle x^{n_j} + \beta^{n_j} / r F'(x^{n_j+1})^*(y - F(x^{n_j})) - x^{n_j+1}, w - x^{n_j+1} \rangle \leq 0 \quad \text{for all } w \in B_R .$$

Therefore we obtain (due to Lemma 8)

$$\limsup_{j \to \infty} \beta^{n_j} / r \langle F'(x^{n_j+1})^*(y - F(x^{n_j})), w - x^{n_j+1} \rangle \leq 0 \quad \text{for all } w \in B_R .$$

To the latter inequality we may add

$$\beta^{n_j} / r \langle (-F'(x^{n_j+1})^* + F'(x^{n_j})^*)(y - F(x^{n_j})), w - x^{n_j+1} \rangle$$

and

$$\beta^{n_j} / r \langle F'(x^{n_j})^*(y - F(x^{n_j})), -x^{n_j} + x^{n_j+1} \rangle$$

resulting in

$$\limsup_{j \to \infty} \beta^{n_j} / r \langle F'(x^{n_j})^*(y - F(x^{n_j})), w - x^{n_j} \rangle \leq 0 \quad \text{for all } w \in B_R , \qquad (3.10)$$

which is possible due to

$$|\langle (-F'(x^{n_j+1})^* + F'(x^{n_j})^*)(y - F(x^{n_j})), w - x^{n_j+1} \rangle| \leq L \|x^{n_j+1} - x^{n_j}\| \|y - F(x^{n_j})\| \|w - x^{n_j+1}\| \overset{j \to \infty}{\longrightarrow} 0$$

and

$$|\langle F'(x^{n_j})^*(y - F(x^{n_j})), -x^{n_j} + x^{n_j+1} \rangle| \leq \sup_{x \in B_R} \|F'(x)^*\| \|y - F(x^{n_j})\| \|x^{n_j} - x^{n_j+1}\| \overset{j \to \infty}{\longrightarrow} 0 .$$

Let us now consider the inner product in (3.10) which we write as

$$\langle F'(x^{n_j})^* y, w - x^{n_j} \rangle - \langle F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle .$$

For the left summand we have by the weak convergence of $\{x^{n_j}\}_{j \in \mathbb{N}}$ or likewise $\{F'(x^\star) x^{n_j}\}_{j \in \mathbb{N}}$ and the assumption of F, $F'(x^{n_j})^* y \overset{j \to \infty}{\longrightarrow} F'(x^\star)^* y$,

$$
\begin{aligned}
\langle F'(x^{n_j})^* y, w - x^{n_j} \rangle &= \langle (F'(x^{n_j})^* - F'(x^\star)^* + F'(x^\star)^*) y, w - x^{n_j} \rangle \\
&= \langle F'(x^{n_j})^* y - F'(x^\star)^* y, w - x^{n_j} \rangle + \langle F'(x^\star)^* y, w - x^{n_j} \rangle \\
&\overset{j \to \infty}{\longrightarrow} \langle F'(x^\star)^* y, w - x^\star \rangle \\
&= \langle F'(x^\star)^*(y - F(x^\star)), w - x^\star \rangle + \langle F'(x^\star)^* F(x^\star), w - x^\star \rangle .
\end{aligned}
$$

14

Therefore (and since $1 \le \beta^{n_j} \le \bar{\beta}$ and again by the weak convergence of $\{x^{n_j}\}_{j \in \mathbb{N}}$), inequality (3.10) transforms to

$$\limsup_{j \to \infty} [\langle F'(x^\star)^*(y - F(x^\star)), w - x^\star \rangle$$
$$+ \langle F'(x^\star)^* F(x^\star), w - x^\star + x^{n_j} - x^{n_j} \rangle - \langle F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle] \le 0$$

$$\Longleftrightarrow$$

$$\limsup_{j \to \infty} [\langle F'(x^\star)^*(y - F(x^\star)), w - x^\star \rangle$$
$$+ \langle F'(x^\star)^* F(x^\star) - F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle] \le 0$$

$$\Longleftrightarrow$$

$$\langle F'(x^\star)^*(y - F(x^\star)), w - x^\star \rangle + \limsup_{j \to \infty} \langle F'(x^\star)^* F(x^\star) - F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle \le 0$$
$$. \tag{3.11}$$

It remains to show that the right summand in (3.11) is for all $w \in B_R$ zero. We have by the assumptions made on F,

$$|\langle F'(x^\star)^* F(x^\star) - F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle|$$
$$= |\langle F'(x^\star)^* F(x^\star) - F'(x^\star)^* F(x^{n_j}) + F'(x^\star)^* F(x^{n_j}) - F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle|$$
$$\le |\langle F'(x^\star)^* F(x^\star) - F'(x^\star)^* F(x^{n_j}), w - x^{n_j} \rangle| + |\langle F'(x^\star)^* F(x^{n_j}) - F'(x^{n_j})^* F(x^{n_j}), w - x^{n_j} \rangle|$$
$$\le \sup_{x \in B_R} \|F'(x)\| \|F(x^\star) - F(x^{n_j})\| \|w - x^{n_j}\| +$$
$$|\langle (F'(x^\star)^* - F'(x^{n_j})^*)(F(x^\star) - F(x^\star) + F(x^{n_j})), w - x^{n_j} \rangle|$$
$$\le \sup_{x \in B_R} \|F'(x)\| \|F(x^\star) - F(x^{n_j})\| \|w - x^{n_j}\| +$$
$$\|(F'(x^\star)^* - F'(x^{n_j})^*) F(x^\star)\| \|w - x^{n_j}\| + L\|x^\star - x^{n_j}\| \|F(x^\star) - F(x^{n_j})\| \|w - x^{n_j}\|$$
$$\overset{j \to \infty}{\longrightarrow} \quad 0 .$$

Consequently, for all $w \in B_R$,

$$\langle F'(x^\star)^*(y - F(x^\star)), w - x^\star \rangle \le 0$$

and the proof is complete. $\qquad\square$

3.3   Strong Convergence

Within this subsection we show that the weak convergence of subsequences towards accumulation points $x^\star$ can be strengthened into convergence in norm topology. Convergence in norm is a very useful property that ensures numerical stability with respect infinite series expansions (i.e. for expansions where Λ is infinite).

Lemma 10 *With the same assumptions as in Proposition 9 and the assumptions* (1.8)-(1.9) *on the nonlinear operator F, there exists a subsequence $\{x^{n'_l}\}_{l\in\mathbb{N}} \subset \{x^n\}_{n\in\mathbb{N}}$ such that $\{x^{n'_l}\}_{l\in\mathbb{N}}$ converges in norm towards the weak accumulation point $x^\star$, i.e.*

$$\lim_{l\to\infty} \|x^{n'_l} - x^\star\| = 0 \ .$$

Proof. Denote by $\{x^{n_j}\}_{j\in\mathbb{N}}$ the subsequence that was introduced in the proof of Proposition 9. Define now $u^j := x^{n_j} - x^\star$, $v^j := x^{n_j+1} - x^\star$, and $\beta^j := \beta^{n_j}$. Due to Lemma 8, we have $\lim_{j\to\infty} \|u^j - v^j\| = 0$. But we also have,

$$
\begin{aligned}
u^j - v^j &= u^j + x^\star - \mathrm{P}_R(u^j + x^\star + \beta^j F'(v^j + x^\star)^*(y - F(u^j + x^\star))) \\
&= u^j + \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star))) \\
&\quad - \mathrm{P}_R(u^j + x^\star + \beta^j F'(v^j + x^\star)^*(y - F(u^j + x^\star))) \\
&= u^j + \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star))) \\
&\quad - \mathrm{P}_R(x^\star + \beta^j F'(v^j + x^\star)^*(y - F(u^j + x^\star)) + u^j) &(3.12) \\
&\quad + \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star)) + u^j) &(3.13) \\
&\quad - \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star)) + u^j) \\
&\quad + \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(u^j + x^\star)) + u^j) &(3.14) \\
&\quad - \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(u^j + x^\star)) + u^j) \ , &(3.15)
\end{aligned}
$$

where we have applied Proposition 9 ($x^\star$ fulfills the necessary condition) and Lemma 4, i.e. $x^\star = \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star)))$. We consider now the sum (3.13)+(3.15), and obtain by the assumptions on F and since the β^j are uniformly bounded,

$$
\begin{aligned}
&\|\mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star)) + u^j) - \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(u^j + x^\star)) + u^j)\| \\
&\leq \|\beta^j F'(x^\star)^*(F(u^j + x^\star) - F(x^\star))\| \leq \bar{\beta} \sup_{x\in B_R} \|F'(x)\| \|F(u^j + x^\star) - F(x^\star)\| \xrightarrow{j\to\infty} 0 \ .
\end{aligned}
$$

The second sum (3.12)+(3.14) yields

$$
\begin{aligned}
&\|\mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(u^j + x^\star)) + u^j) - \mathrm{P}_R(x^\star + \beta^j F'(v^j + x^\star)^*(y - F(u^j + x^\star)) + u^j)\| \\
&\leq \bar{\beta}\left\{\|(F'(x^\star)^* - F'(v^j + x^\star)^*)(y - F(x^\star))\| \right. \\
&\quad \left. + \|(F'(x^\star)^* - F'(v^j + x^\star)^*)(F(x^\star) - F(u^j + x^\star))\|\right\} \\
&\leq \bar{\beta}\left\{\|(F'(x^\star)^* - F'(v^j + x^\star)^*)(y - F(x^\star))\| + L\|v^j\| \|F(x^\star) - F(u^j + x^\star)\|\right\} \xrightarrow{j\to\infty} 0 \ .
\end{aligned}
$$

Consequently, combining $\|u^j - v^j\| \xrightarrow{j\to\infty} 0$ and the two last statements, we observe that

$$\lim_{j\to\infty} \|\mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star)) + u^j) - \mathrm{P}_R(x^\star + \beta^j F'(x^\star)^*(y - F(x^\star))) - u^j\| = 0 \ .$$

The remaining arguments that verify the strong convergence towards zero of a subsequence of u^j are now the same as in [3, Lemma 12]. For the readers convenience, we give all the details.

The sequence β^j is uniformly bounded, therefore there must be at least one accumulation point, which we denote by β^∞. We choose a subsequence $\{j_l\}_{l\in\mathbb{N}}$ such that $\lim_{l\to\infty} \beta^{j_l} = \beta^\infty$. Defining $n'_l := n_{j_l}$, $u'^l := u^{j_l}$, and $v'^l := v^{n_l}$, we have until now

$$\lim_{l\to\infty} \beta^{j_l} = \beta^\infty \ ,$$

$$\lim_{j\to\infty} \|P_R(x^\star + \beta^{j_l} F'(x^\star)^*(y - F(x^\star)) + u'^l) - P_R(x^\star + \beta^{j_l} F'(x^\star)^*(y - F(x^\star))) - u'^l\| = 0 \ .$$

Denote $h^\star := x^\star + \beta^\infty F^*(y - F(x^\star))$ and $h'^l := x^\star + \beta^{j_l} F^*(y - F(x^\star))$. We have now

$$\|P_R(h^\star + u'^l) \quad -P_R(h^\star) - u'^l\|$$
$$\leq \qquad \|P_R(h'^l + u'^l) - P_R(h'^l) - u'^l\|$$
$$\qquad +\|P_R(h'^l + u'^l) - P_R(h^\star + u'^l)\| + \|P_R(h'^l) - P_R(h^\star)\|$$
$$\leq \qquad \|P_R(h'^l + u'^l) - P_R(h'^l) - u'^l\| + 2\|h'^l - h^\star\|$$

Since both terms on the right hand sinde converge to zero for $l \to \infty$, we have

$$\lim_{l\to\infty} \|P_R(h^\star - u'^l) - P_R(h^\star) - u'^l\| = 0 \tag{3.16}$$

Without loss of generality we can assume $\|h^\star\|_1 > R$. By Lemma 2 there exists $\mu > 0$ such that $P_R(h^\star) = S_\mu(h^\star)$. Because $|h^\star_\lambda| \to 0$ as $|\lambda| \to \infty$, this implies that, for some finite $K_1 > 0$, $(P_R(h^\star))_\lambda = 0$ for $|\lambda| > K_1$. Pick now any $\varepsilon > 0$ that satisfies $\varepsilon < \mu/5$. There exist a finite $K_2 > 0$ so that $\sum_{|\lambda|>K_2} |h^\star_\lambda|^2 < \varepsilon^2$. Set $K_0 := \max(K_1, K_2)$, and define the vector $\tilde{h}^\star$ by $\tilde{h}^\star_\lambda = h^\star_\lambda$ if $|\lambda| \leq K_0$, $\tilde{h}^\star_\lambda = 0$ if $|\lambda| > K_0$.
By the weak convergence of the u'^l, we can, for this same K_0, determine $C_1 > 0$ such that, for all $k \geq C_1$, $\sum_{|\lambda|\leq K_0} |u'^l_\lambda|^2 \leq \varepsilon^2$. Define new vectors $\tilde{u}'^l$ by $\tilde{u}'^l_\lambda = 0$ if $|\lambda| \leq K_0$, $\tilde{u}'^l_\lambda = u'^l_\lambda$ if $|\lambda| > K_0$.
Because of 3.16, there exists $C_2 > 0$ such that $\|P_R(h^\star + u'^l) - P_R(h^\star) - u'^l\| \leq \varepsilon$ for $l \geq C_2$. Consider now $l \geq C := \max(C_1, C_2)$. We have

$$\|P_R(\tilde{h}^\star + \tilde{u}'^l) - P_R(\tilde{h}^\star) - \tilde{u}'^l\|$$
$$\leq \|P_R(\tilde{h}^\star + \tilde{u}'^l) - P_R(h^\star + \tilde{u}'^l)\| + \|P_R(h^\star + \tilde{u}'^l) - P_R(h^\star + u'^l)\|$$
$$+\|P_R(h^\star + u'^l) - P_R(h^\star) - u'^l\| + \|P_R(h^\star) - P_R(\tilde{h}^\star)\| + \|u'^l - \tilde{u}'^l\|$$
$$\leq 5\varepsilon$$

On the other hand, Lemma 2 tells us that there exists $\sigma_l > 0$ such that $P_R(\tilde{h}^\star + \tilde{u}'^l) = S_{\sigma_l}(\tilde{h}^\star + \tilde{u}'^l) = S_{\sigma_l}(\tilde{h}^\star) + S_{\sigma_l}(\tilde{u}'^l)$, where we used in the last equality that $\tilde{h}^\star_\lambda = 0$ for $|\lambda| > K_0$ and $\tilde{u}'^l_\lambda = 0$ for $|\lambda| \leq K_0$. From $\|S_\mu(\tilde{h}^\star)\|_1 = R = \|S_{\sigma_l}(\tilde{h}^\star)\|_1 + \|S_{\sigma_l}(\tilde{u}'^C)\|_1$ we conclude that $\sigma_l \geq \mu$ for all $l \geq C$. We then deduce

$$(5\varepsilon)^2 \geq \|P_R(\tilde{h}^\star + \tilde{u}'^l) - P_R(\tilde{h}^\star) - \tilde{u}'^l\|^2$$
$$= \sum_{|\lambda|\leq K_0} |S_{\sigma_l}(\tilde{h}^\star_\lambda) - S_\mu(\tilde{h}^\star_\lambda)|^2 + \sum_{|\lambda|>K_0} |S_{\sigma_l}(\tilde{u}'^l_\lambda) - \tilde{u}'^l_\lambda|^2$$
$$\geq \sum_{|\lambda|>K_0} \left[\max\left(|\tilde{u}'^l_\lambda - \sigma_l|, 0\right) - |\tilde{u}'^l_\lambda|\right]^2$$
$$= \sum_{|\lambda|>K_0} \min\left(|\tilde{u}'^l_\lambda|, \sigma_l\right)^2 \geq \sum_{|\lambda|>K_0} \min\left(|\tilde{u}'^l_\lambda|, \mu\right)^2$$

17

Because we picked $\varepsilon < \mu/5$, this is possible only if $|\tilde{u}_\lambda'^l| \leq \mu$ for all $|\lambda| > K_0$, $l \geq C$, and if , in addition,

$$\left[\sum_{|\lambda| > K_0} |\tilde{u}_\lambda'^l|^2 \right]^{1/2} \leq 5\varepsilon, \ i.e., \ \|\tilde{u}'^l\| \leq 5\varepsilon \tag{3.17}$$

It follows that $\|u'^l\| \leq \|\tilde{u}'^l\| + [\sum_{|\lambda| \leq K_0} |u_\lambda'^l|^2]^{1/2} \leq 6\varepsilon$.

We have thus obtained what we set out to prove: the subsequence $(x^{n_{j_l}})_{l \in \mathbb{N}}$ of $(x^n)_{n \in \mathbb{N}}$ satisfies that, given arbitrary $\varepsilon > 0$, there exists C so that, for $l > C, \|x^{n_{j_l}} - x^\star\| \leq 6\varepsilon$. $\qquad\square$

As mentioned in [3], one can prove at the cost of more technicalities that the whole subsequence $\{x^{n_j}\}_{j \in \mathbb{N}}$ converges in norm towards $x^\star$. We summarize subsections 3.2 and 3.3 in the following proposition.

Proposition 11 *Every weak accumulation point $x^\star$ of the sequence $\{x^n\}_{n \in \mathbb{N}}$ defined by (1.6) fulfills the necessary condition for a minimizer of $\mathcal{D}$ in B_R. Moreover, there exists a subsequence $\{x^{n_j}\}_{j \in \mathbb{N}} \subset \{x^n\}_{n \in \mathbb{N}}$ that converges in norm to $x^\star$.*

3.4 Some Algorithmic Aspects

In the previous subsection we have shown norm convergence for all β^n satisfying Condition (B). This, of course, implies also norm convergence for $\beta^n = 1$ for all $n \in \mathbb{N}$, which corresponds to convergence of the projected classical Landweber iteration. But as we have mentioned above, we intend to accelerate the speed of convergence. Therefore we are interested in choosing, adaptively, larger values for β^n. In particular, by the reasoning made after Definition 1, we like to choose β^n as large as possible. The problem (even for linear operators F) is that the definition of x^{n+1} involves β^n and the inequality (B2) to restrict the choice of β^n uses x^{n+1}. This "implicit" quality does not allow for a straightforward determination of β^n.

Conditions (B1) and (B2) are inspired by classical length-step in the steepest descent algorithm for the unconstrained functional $\|Fx - y\|^2$ (where F is linear) leading to an accelerated Landweber iteration $x^{n+1} = x^n + \gamma^n F^*(y - Fx^n)$, for which γ^n is picked so that it gives a maximal decrease of $\|Fx - y\|^2$, i.e.

$$\gamma^n = \|F^*(y - Fx^n)\|^2 \|FF^*(y - Fx^n)\|^{-2} \ .$$

For nonlinear operators this condition translates into a rather non-practical suggestion for γ^n. In our situation, in which we have to fulfill Condition (B), we may derive a much simpler procedure to find a suitable γ^n (which is in our case β^n/r). Due to Lemma 8 we have monotonicity of $\mathcal{D}$ with respect to the iterates, i.e.

$$L\sqrt{\mathcal{D}(x^n)} \leq L\sqrt{\mathcal{D}(x^{n-1})} \leq \ldots \leq \frac{r}{2} = \max\{ \sup_{x \in B_R} \|F'(x)\|^2, L\sqrt{\mathcal{D}(x^0)}\} \ .$$

Therefore (B3), which was given by

$$L\sqrt{\mathcal{D}(x^n)} \leq \beta^n L\sqrt{\mathcal{D}(x^n)} \leq \frac{r}{2},$$

is indeed a nontrivial condition for $\beta^n \geq 1$. Namely, the larger the decrease of $\mathcal{D}$, the larger we may choose β^n (when only considering (B3)). Condition (B3) can be recast as $1 \leq \beta^n \leq r/(2L\sqrt{\mathcal{D}(x^n)})$ and consequently, by Definition (3.1), an *explicit* (but somewhat "greedy") guess for β^n is given by

$$\beta^n = \max \left\{ \sup_{x \in B_R} \frac{\|F'(x)\|^2}{L\sqrt{\mathcal{D}(x^n)}}, \sqrt{\frac{\mathcal{D}(x^0)}{\mathcal{D}(x^n)}} \right\} \geq 1 \ . \tag{3.18}$$

If this choice fulfills (B2) as well, it is retained; if it does not, it can be gradually decreased (by multiplying it with a factor slightly smaller than 1 until (B2) is satisfied.

As a summary of the above reasoning we suggest the following implementation of the proposed projected steepest descent algorithm.

Projected Steepest Descent Method	
for nonlinear inverse problems	
Given	operator F, its derivative $F'(x)$, data y, some initial guess x^0, and R (sparsity constraint ℓ_1-ball B_R)
Initialization	$r = \max\{2 \sup_{x \in B_R} \|F'(x)\|^2, 2L\sqrt{\mathcal{D}(x^0)}\}$, set $q = 0.9$ (as an example)

Iteration	**for** $n = 0, 1, 2, \ldots$ **until** a preassigned precision / maximum number of iterations
	
	 1. $\beta^n = \max \left\{ \sup_{x \in B_R} \frac{\|F'(x)\|^2}{L\sqrt{\mathcal{D}(x^n)}}, \sqrt{\frac{\mathcal{D}(x^0)}{\mathcal{D}(x^n)}} \right\}$
	
	 2. $x^{n+1} = \mathrm{P}_R \left(x^n + \frac{\beta^n}{r} F'(x^{n+1})^*(y - F(x^n)) \right)$; by fixed point iteration
	
	 3. verify (B2): $\beta^n \|F(x^{n+1}) - F(x^n)\|^2 \leq \frac{r}{2}\|x^{n+1} - x^n\|^2$
	
	     if (B2) is satisfied increase n and go to **1.**
	     otherwise set $\beta^n = q \cdot \beta^n$ and go to **2.**
	
	 end

4   Numerical Experiment: A Nonlinear Sensing Problem

Our numerical experiment centers around a nonlinear sampling problem that is very closely related to the sensing problem considered in [13]. The authors of [13] have studied a sensing setup in which a continuous-time signal is mapped by a memoryless, invertible and nonlinear transformation, and then sampled in a non-ideal manner. Such scenarios may appear in acquisition systems where the sensor introduces static nonlinearities, before the signal is sampled

by a usual analog-to-digital converter. In [13] a theory and an algorithm is developed that allow a perfect recovery of a signal within a subspace from its nonlinear and non-ideal samples. In our setup we drop the invertibility requirement of the nonlinear transformation, which is indeed quite restrictive. Moreover, we focus on a subclass of problems in which the signal to be recovered is supposed to have sparse expansion.

Let us specify the sensing model. Assume we are given a reconstruction space $\mathcal{A} \subset X$ (e.g. $L_2(\mathbb{R})$) which is spanned by $\Psi = \{a_\lambda : \lambda \in \Lambda\}$ and where the family Ψ forms a frame for $\mathcal{A}$ with frame bounds $0 < A_1 \leq A_2 < \infty$. With Ψ we associate two mappings, the analysis and synthesis operator,

$$A : \mathcal{A} \ni f \mapsto \{\langle f, a_\lambda \rangle\}_{\lambda \in \Lambda} \in \ell_2(\Lambda) \ \text{ and } \ A^* : \ell_2(\Lambda) \ni x \mapsto \sum_{\lambda \in \Lambda} x_\lambda a_\lambda \ .$$

We assume that the function/signal f we wish to recover has a sparse expansion in $\mathcal{A}$. The sensing model is now determined by the nonlinear transformation $M : \mathcal{A} \to Y$ of the continuous-time function f that is point-wise given by the regularized modulus function (to have some concrete example for the nonlinear transformation)

$$M : f \mapsto M(f) = |f|_\varepsilon := \sqrt{f^2 + \varepsilon^2} \ .$$

This nonlinearly transformed f is then sampled in a possibly non-ideal fashion by some sampling function s yielding the following sequence of samples,

$$SM(f) = \{\langle s(\cdot - nT), M(f) \rangle_Y\}_{n \in \mathbb{Z}}.$$

As for Ψ, we assume that the family $\Sigma = \{s(\cdot - nT_s), n \in \mathbb{Z}\}$ forms a frame with bounds $0 < S_1 \leq S_2 < \infty$. The goal is to reconstruct f from its samples $y = (S \circ M)(f)$. Since f belongs to $\mathcal{A}$, the reconstruction of f is equivalent with finding a sequence x such that $y = (S \circ M \circ A^*)(x)$. If Ψ forms a basis the searched for sequence $x \in \ell_2(\Lambda)$ is unique; otherwise there might be several different sequences leading to the same function f. Among all possible solutions, we aim (as mentioned above) to find those sequences that have small ℓ_1 norm. As y might be not directly accessible (due to the presence of measurement noise) and due to the nonlinearity of the operator M, it seems more practical not to solve $y = (S \circ M \circ A^*)(x)$ directly, but to find an approximation $\hat{x}$ such that

$$\hat{x} = \arg \min_x \|F(x) - y\|^2 \ \text{ and } \ \hat{x} \in B_R \ ,$$

where we have used the shorthand notation $F := S \circ M \circ A^*$ and where the ℓ_1 ball B_R restricts x to have a certain preassigned sparsity.

In order to apply our proposed accelerated steepest descent iteration,

$$x^{n+1} = \mathrm{P}_R \left(x^n + \frac{\beta^n}{r} F'(x^{n+1})^*(y - F(x^n)) \right),$$

to derive an approximation to $\hat{x}$, we have to determine the constants r, see(3.1), and the Lipschitz constant L. This requires a specification of Ψ and Σ. One technically motivated

choice in signal sampling is the cardinal sine function (here used to generate Ψ as well as Σ). This function can be defined as the inverse Fourier transform of the characteristic function of the frequency interval $[-\pi, \pi]$, i.e.

$$\sqrt{2\pi}\mathrm{sinc}(\pi t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \chi_{[-\pi,\pi]}(\omega)e^{it\omega}d\omega \ .$$

Therefore, the resulting function spaces are spaces of band limited functions. The inverse Fourier transform of the L_2 normalized characteristic function $\frac{1}{\sqrt{2\Omega}}\chi_{[-\Omega,\Omega]}$ yields

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\Omega}}\chi_{[-\Omega,\Omega]}(\omega)e^{it\omega}d\omega = \sqrt{\frac{\Omega}{\pi}}\,\mathrm{sinc}(\Omega t)$$

leading to the following definition of L_2 normalized and translated cardinal sine functions,

$$a_n(t) \ = \ \frac{1}{\sqrt{D_a}}\mathrm{sinc}\left(\frac{\pi}{D_a}(t - nT_a)\right) \ , \quad \text{i.e. } \Omega = \frac{\pi}{D_a} \quad \text{and} \tag{4.1}$$

$$s_n(t) \ = \ \frac{1}{\sqrt{D_s}}\mathrm{sinc}\left(\frac{\pi}{D_s}(t - nT_s)\right) \ , \quad \text{i.e. } \Omega = \frac{\pi}{D_s} \tag{4.2}$$

that determine Ψ and Σ. The parameters D_a and D_s are fixed and specify here the frequency cut off, whereas T_a and T_s fix the time step sizes. For all $n \in \mathbb{Z}$ we have $\|a_n\|_2 = \|s_n\|_2 = 1$. Moreover, it can be easily retrieved that

$$\langle a_n, a_m \rangle = \mathrm{sinc}\left(\frac{\pi}{D_a}(n - m)T_a\right) \quad \text{and} \quad \langle s_n, s_m \rangle = \mathrm{sinc}\left(\frac{\pi}{D_s}(n - m)T_s\right) \ . \tag{4.3}$$

As long as $T_a/D_a, T_s/D_s \in \mathbb{Z}$, the families Ψ and Σ form orthonormal systems. The inner products (4.3) are the entries of the Gramian matrices AA^* and SS^*, respectively, for which we have $\|AA^*\| = \|A\|^2 = \|A^*\|^2 \leq A_2$ and $\|SS^*\| = \|S\|^2 = \|S^*\|^2 \leq S_2$.

Let us now determine r and L. To this end we have to estimate $\sup_{x \in B_R} \|F'(x)\|^2$. For given $x \in B_R$, it follows that

$$\|F'(x)\| \ = \ \sup_{h \in \ell_2, \|h\|=1} \|F'(x)h\| = \|SM'(A^*x)A^*h\| \leq \|S\|\,\|M'(A^*x)\|\,\|A^*\| \ . \tag{4.4}$$

Moreover, due to (4.1),

$$\|M'(A^*x)\|^2 \ = \ \sup_{h \in \Lambda_2, \|h\|=1} \|M'(A^*x)h\|^2 = \int_{\mathbb{R}} |(A^*x)(t)|^2|((A^*x)(t))^2 + \varepsilon^2|^{-1}|h(t)|^2 dt$$

$$\leq \ \frac{1}{\varepsilon^2} \int_{\mathbb{R}} (\sum_n |x_n||a_n(t)|)^2 |h(t)|^2 dt \leq \frac{\|x\|_1^2}{\varepsilon^2 D_a} \ .$$

Therefore, we finally obtain

$$\sup_{x \in B_R} \|F'(x)\|^2 \leq \|S\|^2\,\|A^*\|^2\,\frac{R^2}{\varepsilon^2 D_a} \leq S_2\,A_2\,\frac{R^2}{\varepsilon^2 D_a} \ . \tag{4.5}$$
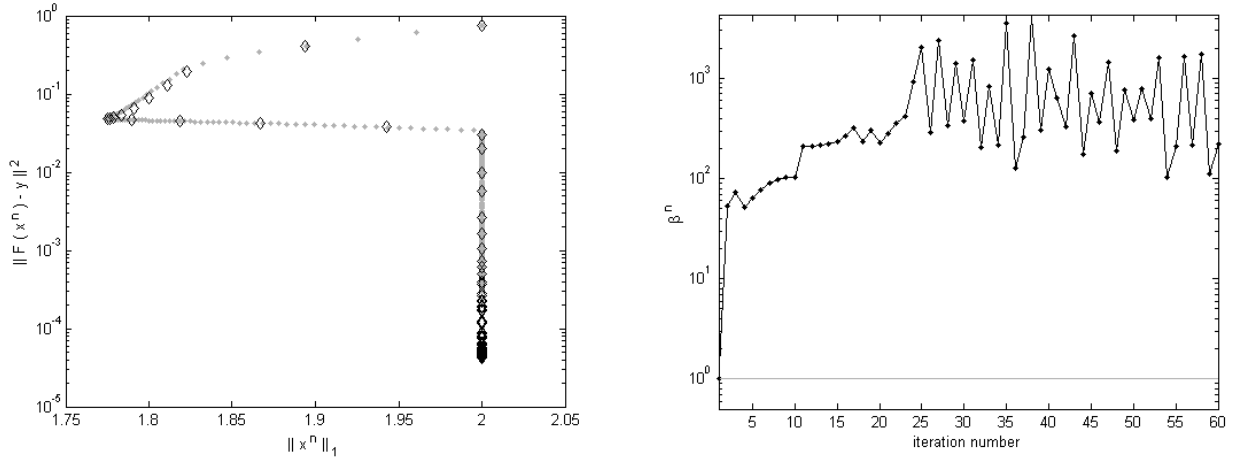
Figure 1: The left image shows the sparsity to residual plot. The black diamonds correspond to the accelerated iteration. For the non-accelerated iteration we have plotted every 20th iteration (gray dots). The right image visualizes the sequence of β^n (black) for the accelerated iteration. The gray line corresponds to $\beta = 1$.

The Lipschitz continuity of F' is characterized by $\|F'(\tilde{x}) - F'(x)\| \leq L\|\tilde{x} - x\|$, for all $x, \tilde{x} \in B_R$. In order to find the Lipschitz constant L, we directly derive

$$
\begin{aligned}
\|F'(\tilde{x}) - F'(x)\| &= \sup_{h \in \ell_2, \|h\|=1} \|F'(\tilde{x})h - F'(x)h\| \\
&= \sup_{h \in \ell_2, \|h\|=1} \|SM'(A^*\tilde{x})A^*h - SM'(A^*x)A^*h\| \\
&\leq \|S\| \, \|M'(A^*\tilde{x}) - M'(A^*x)\| \, \|A^*\| \,,
\end{aligned} \tag{4.6}
$$

and with $M''(f) = \varepsilon^2(f^2 + \varepsilon^2)^{-3/2}$ it follows

$$
\begin{aligned}
\|M'(A^*\tilde{x}) - M'(A^*x)\|^2 &= \sup_{h \in L_2, \|h\|=1} \int_{\mathbb{R}} |M'(A^*\tilde{x}(t)) - M'(A^*x(t))|^2 |h(t)|^2 dt \\
&\leq \sup_{h \in L_2, \|h\|=1} \int_{\mathbb{R}} \frac{1}{\varepsilon^2} |A^*\tilde{x}_n(t) - A^*x(t)|^2 \, |h(t)|^2 dt \\
&\leq \sup_{h \in L_2, \|h\|=1} \int_{\mathbb{R}} \frac{1}{\varepsilon^2} \left(\sum_{n \in \mathbb{Z}} |(\tilde{x}_n - x_n)| \, |a_n(t)| \right)^2 |h(t)|^2 dt \\
&\leq \sup_{h \in L_2, \|h\|=1} \int_{\mathbb{R}} \sum_{n \in \mathbb{Z}} |a_n(t)|^2 \, |h(t)|^2 dt \, \frac{1}{\varepsilon^2} \|\tilde{x} - x\|^2 \,.
\end{aligned}
$$

To finally bound the last quantity, we have to estimate $\sum_{n \in \mathbb{Z}} |a_n(t)|^2$ independently on $t \in \mathbb{R}$. With definition (4.1), we observe that

$$
\sum_{n \in \mathbb{Z}} |a_n(t)|^2 = \frac{1}{D_a} \sum_{n \in \mathbb{Z}} \operatorname{sinc}^2 \left(\frac{\pi}{D_a} t - n \frac{\pi T_a}{D_a} \right) \tag{4.7}
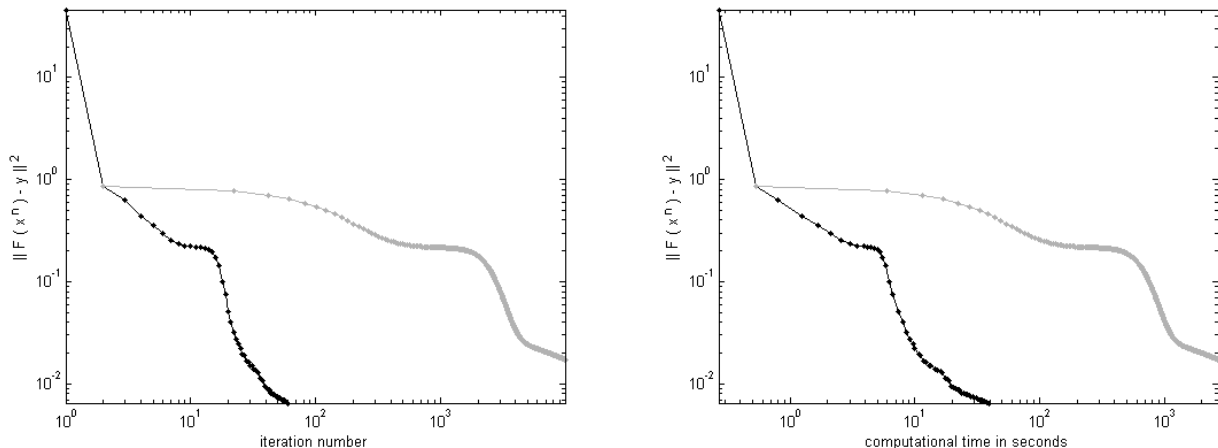$$

22

Figure 2: These images represent the residual evolution with respect to the number of iterations (left) and the computational time (right). The black dotted curves represent the residual evolution for the accelerated and the gray dotted curves for the non-accelerated scheme.

is a periodic function with period T_a. Therefore it is sufficient to analyze (4.7) for $t \in [0, T_a]$. The sum in (4.7) is maximal for $t = 0$ and $t = T_a$. Consequently, with

$$\sum_{n \in \mathbb{Z}} \operatorname{sinc}^2\left(n\frac{\pi T_a}{D_a}\right) = 1 + \sum_{n \in \mathbb{Z} \setminus \{0\}} \operatorname{sinc}^2\left(n\frac{\pi T_a}{D_a}\right) \leq 1 + \frac{2 D_a^2}{\pi^2 T_a^2} \sum_{n \in \mathbb{N} \setminus \{0\}} \frac{1}{n^2} = 1 + \frac{4 D_a^2}{\pi^2 T_a^2}$$

we obtain by combining (4.6) and (4.7),

$$\|F'(\tilde{x}) - F'(x)\| \leq L \|\tilde{x} - x\| , \quad \text{with} \quad L := \frac{1}{\varepsilon} \sqrt{\frac{1}{D_a} + \frac{4 D_a}{\pi^2 T_a^2}} \sqrt{S_2} \sqrt{A_2} . \tag{4.8}$$

In our concrete example (visualized in Figure 3) the ansatz space $\mathcal{A} \subset L_2(\mathbb{R})$ is spanned by functions a_n with $D_a = 0.4$ and time step size $T_a = 0.1$. The sampling map S is determined by $T_s = 0.2$ and $D_s = 0.1$. The synthetic signal which we aim to reconstruct is given by

$$f(t) = a_{-2}(t) - 0.5 a_{2.5}(t) .$$

For the numerical implementation we have restricted the computations to the finite interval $[-10, 10]$ which was discretized by the grid $t_k = -10 + 0.05\, k$ with $k = 0, 1, 2, \ldots$. The bounds A_2 and S_2 are estimated by the eigenvalues of adequately corresponding finite dimensional approximations of the Gramian matrices $\langle a_n, a_m \rangle$ and $\langle s_n, s_m \rangle$. For the radius of the ℓ_1 ball (determined the sparsity constraint) we have picked $R = 2$. Of course, this seems to be an arbitrary guess but it just includes some a-priori knowledge of f. In Figure 1 (left diagram) one can observe that the iterates live (from a certain number of iterations on) on the boundary of the ℓ_1 ball. As also discussed in [3], partly better results can be obtained when slowly increasing the radius, i.e.

$$R^n = (n + 1)R/N ,$$

23

where n is the iteration index and N stands for a prescribed number of iterations. In Figure 1 (right image) one finds that β^n varies significantly from one to another iteration. This verifies the usefulness of Condition (B). From the first iteration on, the values for β^n are obviously larger than one and grow in the first phase of the iteration process (for the accelerated method only the first 60 iterations are shown). But the main impact manifests itself more in the second half of the iteration ($n > 20$) where the non-accelerated variant has a much less decay of $\sqrt{\mathcal{D}(x^n)}$, see Figure 2. There the values of β^n vary around 10^3 and allow that impressive fast and rapid decay of $\sqrt{\mathcal{D}(x^n)}$ of the accelerated descent method. For the non-accelerated method we had to compute 10^4 iterations to achieve reasonable small residuals $\sqrt{\mathcal{D}(x^n)}$ (but even then being far off the nice results achieved by the accelerated scheme). The right plot in Figure 2 sketches the residual decay with respect to the overall computational time that was practically necessary. Both curves (the black and the gray) were of course obtained on the same machine under same conditions. The achieved time reduction is remarkable as the accelerated iteration method has required many additional loops of the individual fixed point iterations in order to find the optimal β^n. In particular, the final residual value after $n = 10.000$ iterations for the non-accelerated method was $\sqrt{\mathcal{D}(x^{10000})} = 0.0172$. This value was reached by the accelerated method after $n = 28$ iteration steps (the final value after $n = 60$ iterations was $\sqrt{\mathcal{D}(x^{60})} = 0.0065$). The overall computational time consumption of the non-accelerated method to arrive at $\sqrt{\mathcal{D}(x^{10000})} = 0.0172$ was 45min and 2s, whereas the time consumption for the accelerated method for the same residual discrepancy was only 11.8s, i.e. 229 times faster. The finally resulting reconstruction including a diagram showing the nonlinearly sampled data is given in Figure 3.

Summarizing this numerical experiment, we can conclude that all the theoretical statements of the previous sections can be verified. For this particular nonlinear sensing problem we can achieve an impressive factor of acceleration. But this, however, holds for this concrete setting. There is no proved guaranty that the same can be achieved for other applications. Nevertheless, experiments in the field of medical imaging show a very similar behavior.

References

[1] K. Bredies, D.A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Application*, 42(2):173–193, 2009.

[2] I. Daubechies, M. Defrise, and C. DeMol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math*, 57:1413–1541, 2004.

[3] I. Daubechies, M. Fornasier, and I. Loris. Accelerated projected gradient methods for linear inverse problems with sparsity constraints. *J. Fourier Anal. Appl.*, to appear, 2008.

[4] I. Daubechies, G. Teschke, and L. Vese. Iteratively solving linear inverse problems with general convex constraints. *Inverse Problems and Imaging*, 1(1):29–46, 2007.

[5] I. Daubechies, G. Teschke, and L. Vese. On some iterative concepts for image restoration. *Advances in Imaging and Electron Physics*, 150:1–51, 2008.

A-atom with LPF π/D_a (black, $D_a=0.4$), S-atom with LPF π/D_s (gray, $D_s=0.2$)

Simulated function f

f (black), S(f) (dots), S*S(f) (gray)

f (gray), f_{rec} after 60 iterations (black) with sampling widths Ts=0.1 Ta=0.1
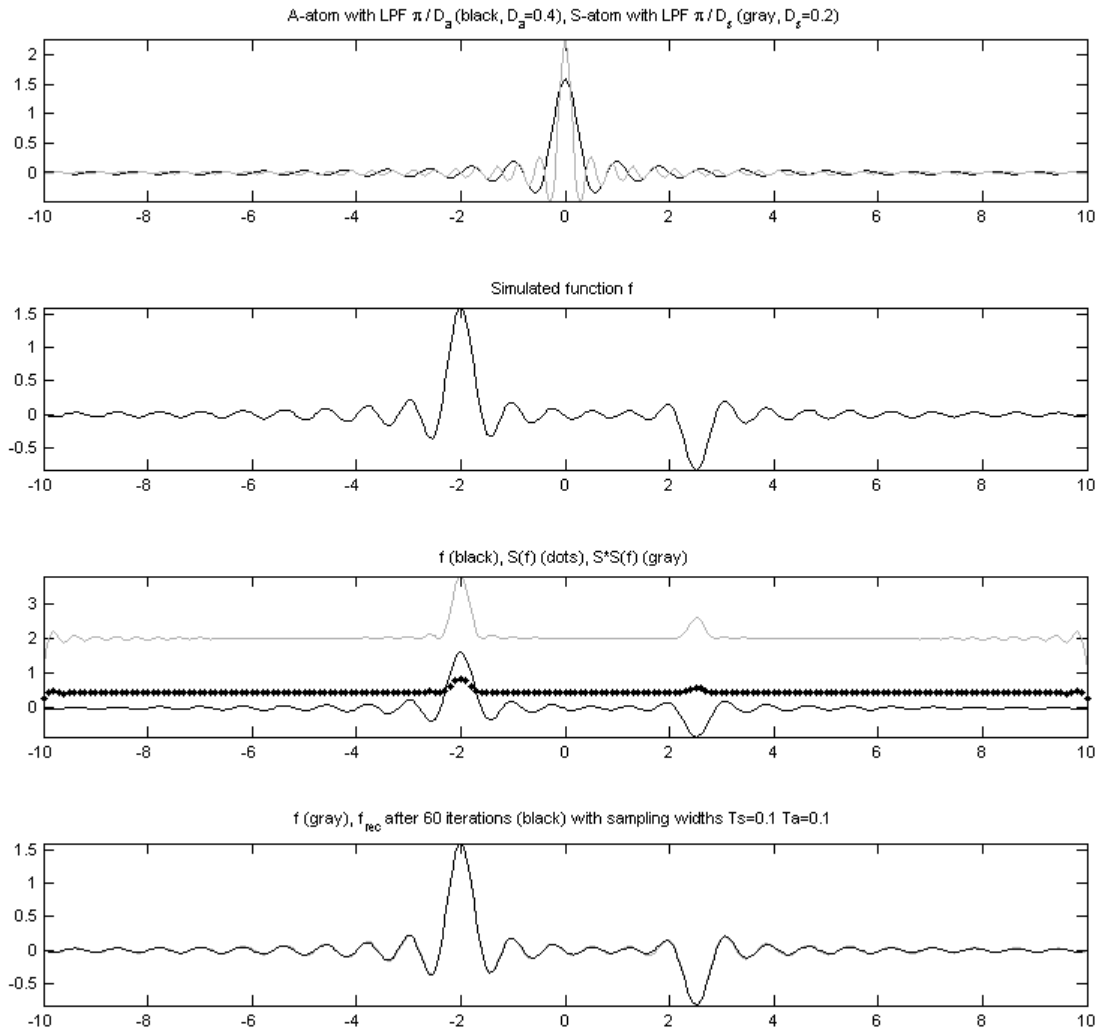
Figure 3: This overview plot shows the used atoms a_0 and s_0 (1st row), the simulated signal (2nd row), the nonlinearly and non-ideally sampled values (3rd row), and the final approximation $A^*x^{60} \in \mathcal{A}$ that was computed with accelerated iteration scheme.

[6] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1996.

[7] M. Fornasier. Domain decomposition methods for linear inverse problems with sparsity constraints. *Inverse Problems*, 23:2505, 2007.

[8] M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraint. *Preprint*, 2006.

[9] R.D. Nowak M.A.T. Figueiredo and S.J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.*, 1(4), 2007.

[10] R. Ramlau and G. Teschke. Tikhonov Replacement Functionals for Iteratively Solving Nonlinear Operator Equations. *Inverse Problems*, 21:1571–1592, 2005.

[11] R. Ramlau and G. Teschke. A Projection Iteration for Nonlinear Operator Equations with Sparsity Constraints. *Numerische Mathematik*, 104:177–203, 2006.

[12] G. Teschke. Multi-Frame Representations in Linear Inverse Problems with Mixed Multi-Constraints. *Applied and Computational Harmonic Analysis*, 22:43 – 60, 2007.

[13] Y. Eldar T.G. Dvorkind and E. Matusiak. Nonlinear and Non-Ideal Sampling: Theory and Methods. *IEEE Trans. on Signal Processing*, 56(12), 2008.

[14] E. van den Berg and M.P. Friedlander. In pursiut of a root. *Preprint*, 2007.