On the convergence speed of iterative methods for linear inverse problems with sparsity constraints

D A Lorenz\(^1\) and K Bredies\(^1\)

\(^1\)Fachbereich 3, Mathematik und Informatik, Universität Bremen, Bibliothekstrasse 1, 28359 Bremen, Germany

E-mail: dlorenz@math.uni-bremen.de

Abstract. In this article we report on different iterative algorithms for the minimization of Tikhonov type functionals which involve sparsity constraints in form of \(\ell^p\)-penalties. We summarize results on the well known iterated soft thresholding, the iterated hard thresholding and a semismooth Newton approach. While the first two converge globally but very slow, the last one converges only locally but superlinearly. Furthermore we propose a combination of soft and hard thresholding. While this method has theoretically the same convergence rate than the soft thresholding (namely it converges linearly), our numerical experiments show that the combined approach is almost always better than both methods alone.

1. Introduction

Recently sparsity constraints have gained a lot of attention in different fields of mathematics. Due to the influential paper [4] sparsity constraints are by now a very popular tool for the regularization of inverse problems, see e.g. [14, 18, 6, 1]. The most popular algorithm for the minimization of regularizing functionals with sparsity constraints is the iterated soft thresholding algorithm which was introduced in [7, 17]. It is now a folk wisdom that this algorithm converges very slowly.

The aim of this article is twofold. First we outline different approaches to the minimization of regularizing functionals with sparsity constraints by means of generalized gradient methods and a semismooth Newton approach and analyze the convergence behavior of the obtained algorithms. Second we present different numerical experiments that illustrate the convergence speed of the different methods.

To fix notation for this article let \(K : \ell^2 \to \mathcal{H}\) be a linear and bounded operator mapping the sequence space \(\ell^2\) to some Hilbert space \(\mathcal{H}\). The regularizing functional we consider is

\[
\Psi(u) = \frac{\|Ku - f\|^2}{2} + \sum_k w_k |u_k|^p
\]  

where \(f \in \mathcal{H}\), \(w_k \geq w_0 > 0\) and \(1 \leq p \leq 2\). The sequence \(w = (w_k)\) plays the role of the regularization parameter and regularizes each component of \(u\) independently.

The problem above may be motivated as follows. We are given a linear operator equation \(Au = f\) with a linear and bounded operator \(A : \mathcal{H}' \to \mathcal{H}\) and possibly noisy data \(f\). Often the inversion of \(A\) is ill-posed and practically impossible. Hence regularization is required. A special
class of regularization methods is, as motivated in [4], the regularization by sparsity constraints. Let \((\psi_k)_k\) be an orthonormal basis of \(H'\) and assume that the unknown solution \(u^*\) is composed of a few components in this basis. A way to use this knowledge as a regularization is to minimize the functional
\[
\frac{\|Au - f\|^2}{2} + \sum_k w_k |\langle u, \psi_k \rangle|^p.
\]
With the help of the synthesis operator \(B : \ell^2 \rightarrow H'\) given by \(Bc = \sum_k c_k \psi_k\) and the operator \(K = AB\) we can rewrite this precisely as (1).

The outline of the paper is as follows. The Section 2 introduces the generalized gradient methods and show how they can be applied to the minimization of the functional (1). Section 3 shows briefly how Newton methods may be used and Section 4 treats the problem how to check if an algorithm for the minimization of (1) has converged without knowing the true minimizer. Section 5 presents numerical experiments on the different methods and the last section gives a short conclusion.

2. Generalized gradient methods

This section introduces generalized gradient methods for the minimization of functionals of type (1). These methods can be motivated as follows. In the theory of optimization an important class of problems are constrained minimization problems
\[
\min_{u \in \Omega} F(u) \quad (2)
\]
where \(\Omega\) is a convex and closed subset of a Hilbert space \(H\) and \(F : H \rightarrow R\). Two well known algorithms for the solution of these problems are the gradient projection method and the conditional gradient method. Both methods will be described in the following subsections. Since the idea for their generalization is the same for both of them, we give it now: The constrained minimization problem (2) can be rephrased as
\[
\min_{u \in H} F(u) + I_\Omega(u) \quad (3)
\]
where \(I_\Omega\) is the indicator function of the set \(\Omega\), i.e. it is zero for \(u \in \Omega\) and infinity otherwise. Note that \(I_\Omega\) is a convex and lower semicontinuous functional (the latter since \(\Omega\) is closed) and hence we can consider a generalization of constrained minimization problems by
\[
\min_{u \in H} F(u) + \Phi(u) \quad (4)
\]
for a general convex and lower semicontinuous functional \(\Phi\).

2.1. The gradient projection method and its generalization

Assume that the functional \(F\) is differentiable and that \(F'\) is Lipschitz continuous with constant \(L\). The gradient projection method for the minimization of (3) reads as follows:

(i) Choose an initial value \(u^0 \in \Omega\) and a stepsize \(0 < s < 2/L\).
(ii) Update by the iteration
\[
u^{n+1}_\Omega = P_\Omega(u^n - sF'(u^n))
\]
where \(P_\Omega\) denotes the orthogonal projection onto the set \(\Omega\).
Hence, the gradient projection method does a gradient descent step and projects the result back onto Ω.

To motivate the generalization we make the observation that the projection onto Ω is characterized by

\[ P_{Ω}(u) = \arg\min_v \frac{\|v - u\|^2}{2} + I_{Ω}(v) \]

If we now deal with the problem (4) we introduce the mapping

\[ P_s(u) = \arg\min_v \frac{\|v - u\|^2}{2} + s\Phi(v) \]

as a generalized projection (where we omitted the dependence on Φ). Note that the mapping \( P_s \) is also called proximal mapping in the context of variational analysis [15]. We state the generalized gradient projection method as:

(i) Choose an initial value \( u_0 \) such that \( \Phi(u_0) < \infty \) and a stepsize \( 0 < s < 2/L \).

(ii) Update by the iteration

\[ u^{n+1} = P_s(u^n - sF'(u^n)) \]

Gradient projection methods have been used for minimization problems with sparsity constraints also in [5, 8].

We now turn to the concrete problem of minimizing (1), i.e. in this case we have \( F(u) = \|Ku - f\|^2/2 \) and hence \( F'(u) = K^*(Ku - f) \) and \( L \leq \|K\|^2 \). The functional \( \Phi(u) = \sum_k w_k|u_k|^p \) is known to be convex and lower semicontinuous. The corresponding generalized projection is

\[ P_s(u) = \arg\min_v \frac{\|v - u\|^2}{2} + s\sum_k w_k|u_k|^p \]

and is known to be a shrinkage operator. It is given by

\[ P_s(u)_k = S_{sw,p}(u)_k = S_{sw_k,p}(u_k) \] (5)

where the component functions \( S_{w,p} \) are

\[ S_{w,p}(x) = \begin{cases} \frac{\text{sgn}(x)\max(|x| - w, 0)}{1 + 2w}, & p = 1 \\ G_{w,p}^{-1}(x), & 1 < p < 2 \\ x, & p = 2 \end{cases} \] (6)

where \( G_{w,p}(y) = y + wp\text{sgn}(y)|y|^{p-1} \), see [4, 12] for detailed explanations.

Hence, the generalized projected gradient method applied to the minimization of (1) now resembles the iterated soft thresholding:

(i) Choose an initial value \( u^0 \) such that \( \Phi(u^0) < \infty \) and a stepsize \( 0 < s < 2/\|K\|^2 \).

(ii) Update by the iteration

\[ u^{n+1} = S_{sw,p}(u^n - sK^*(Ku^n - f)) \]

Note that the stepsize \( s = 1 \) as in [4] is ok as soon as \( \|K\| < \sqrt{2} \) which is a little weaker than \( \|K\| < 1 \) as needed in [4].

The paper [2] shows that the iterated soft-thresholding with \( p = 1 \) converges with linear speed towards the minimizer as soon as the operator \( K \) is injective. This means that there is
a constant $C > 0$ and a $\lambda \in [0, 1]$ such that for the iterates $u^n$ the iterated soft-thresholding and the minimizer $u^*$ it holds
\[ \|u^n - u^*\| \leq C\lambda n. \]
This is a remarkable result since in the case $p = 1$ the iteration is only non-expansive and not contractive (i.e. the Lipschitz constant is exactly one) and linear convergence can not be obtained by Banach-like arguments.

As argued in [2] the constant $\lambda$ is typically very close to one and hence the convergence is in practice very slow. For non-injective $K$ similar results are expected. The paper [5] deals with the possibility to speed up the iteration be choosing larger stepsizes while [8] uses Barzilai-Borwein stepsizes and a backtracking line search.

2.2. The conditional gradient method and its generalization

The conditional gradient method for problem (3) reads as follows:
(i) Choose an initial value $u^0 \in \Omega$.
(ii) Calculate a descent direction
\[ v^n = \arg\min_{v \in \Omega} \langle F'(u^n), v \rangle \]
(iii) Choose a stepsize $s_n$ and update
\[ u^{n+1} = u^n + s_n(v^n - u^n). \]
Here the generalization to problem (4) is straightforward: We calculate the descent direction as
\[ v^n = \arg\min_{v \in \mathcal{H}} \langle F'(u^n), v \rangle + \Phi(v). \]
Note that $v^n$ can be written as $v^n = (\partial \Phi)^{-1}(-F'(u^n))$ and hence we need that the subgradient $\partial \Phi$ is invertible. To apply this algorithm to the problem of minimizing $\Psi$ from (1) we have to do a modification because the subgradient of $\Phi(u) = \sum_k w_k |u_k|^p$ is not onto for $p = 1$. There are at least two different possibilities. The first one is described in [3] and uses the splitting
\[ F(u) = \frac{\|Ku - f\|^2}{2} - \frac{\lambda \|u\|^2}{2}, \quad \Phi(u) = \frac{\lambda \|u\|^2}{2} + \sum_k w_k |u_k|^p \]
for some $\lambda > 0$.

As shown in [3] this again leads to the iterated soft thresholding. Note that this case one has to deal with a non-convex $F$.

The second possibility is described in [1] and uses the following observation. Let $u^*$ be a minimizer of (1). Then one can estimate
\[ \|u^*\|^p \leq \frac{\|f\|^2}{2w_0} \]
and hence, one has to a-priori bound
\[ |u_k^*| \leq \left( \frac{\|f\|^2}{2w_0} \right)^{1/p} =: S_0 \]
for the coefficients of the minimizer. Consequently the minimizer of the functional does not change if we change the penalty term for values which are larger than $S_0$. We do this by introducing the function
\[ \varphi_p(x) = \begin{cases} |x|^p & \text{for } |x| \leq S_0 \\ \frac{p}{2S_0^p} x^2 + \left(\frac{2}{p} - 1\right)S_0^2 & \text{for } |x| > S_0 \end{cases} \]
(7)
The function $\varphi_p$ is a $C^1$ quadratic extension of the $p$-th power. We now consider the functional

$$\tilde{\Psi}(u) = \frac{\|Ku - f\|^2}{2} + \sum_k w_k \varphi_p(u_k)$$

(8)

which has the same minimizers as $\Psi$ from (1). It can be shown that the splitting

$$F(u) = \frac{\|Ku - f\|^2}{2}, \quad \Phi(u) = \sum_k w_k \varphi_p(u_k).$$

has all the properties that ensure convergence of the generalized conditional gradient method, again, see [1]. It is an easy calculation to show that the descent direction $v^n$ has the form

$$v^n_k = H_{\omega,p,\gamma}(-K^*(Ku - f))_k = H_p\left(\frac{-(K^*(Ku - f))_k}{w_k}\right)$$

with the function

$$H_p(x) = \begin{cases} \left(\frac{|x|}{p}\right)^{1/(p-1)} \sgn(x) & \text{for } |x| \leq pS_0^{p-1} \\ \frac{p}{S_0^{p-1}} \left(\frac{S_0^{p-1}}{x}\right)^{p-1} & \text{for } |x| > pS_0^{p-1} \end{cases}$$

where we formally set

$$|x|^{1/p} = \begin{cases} 0 & \text{if } |x| \leq 1 \\ \infty & \text{if } |x| > 1. \end{cases}$$

Note that for $p = 1$ the function $H_1$ is the hard-shrinkage function. In [1] it is shown that the stepsize

$$s_n = \min \left\{ 1, \sum_{k=1}^\infty w_k \left( \varphi_p(u^n_k) - \varphi_p(v^n_k) + (K^*(Ku^n - f))_k(u^n_k - v^n_k) \right) / \|K(u^n - u^n)\|^2 \right\}$$

guarantees convergence of the algorithm. Moreover it is shown in [1] that the iterated hard-thresholding converges linear for $1 < p \leq 2$ and like

$$\|u^n - u^*\| \leq Cn^{-1/2}$$

for $p = 1$. This is theoretically slower than the iterated soft-thresholding but in practice the hard-thresholding is often faster especially when the sparsity is high.

3. Higher order methods

The previous section showed that iterated soft as well as hard thresholding are gradient type algorithms and hence a slow convergence can be expected. In this section the applicability of higher order methods for the minimization of (1) is discussed. We restrict ourself to the case $p = 1$. The minimizer $u^*$ of $\Psi = F + \Phi$ is characterized by

$$-F'(u^*) \in \partial \Phi(u^*).$$

(9)

Since the subgradient $\partial \Phi$ is a maximal monotone operator we can reformulate this characterization for every $\gamma > 0$ as

$$u^* = (I + \gamma \partial \Phi)^{-1}(u^* - \gamma F'(u^*)).$$

(10)
The advantage of (10) over (9) is that it is an equation and not an inclusion and furthermore it may be “smoother” as will be seen below.

In the special case

\[ F(u) = \frac{\| Ku - f \|^2}{2}, \quad \Phi(u) = \sum w_k |u_k| \]

the inclusion (9) looks like

\[ -(K^*(Ku^* - f))_k \in w_k \text{sgn}(u_k^2) \]

while the equation (10) is

\[ u^* = S_{\gamma w,1}(u^n - \gamma K^*(Ku^n - f)). \quad (11) \]

The function \( S_{\gamma w,1} \) defined by (5) and (6) is the well known soft thresholding operator. The paper [9] analyzes a semismooth Newton method for the solution of the equation (10) and hence, minimizing (1).

It turns out that the semismooth Newton algorithm is an active set strategy (which is not surprising, see [11]) and has the following form:

(i) Initialize \( u^0 \) and choose \( \gamma > 0 \) and set \( n = 0 \).

(ii) Compute the soft-thresholding

\[ v^n = S_{\gamma w,1}(u^n - \gamma K^*(Ku^n - f)). \]

(iii) Calculate the active and the inactive set:

\[ A_n = \{ k \in \mathbb{N} \mid v^n_k \neq 0 \} \]
\[ I_n = \{ k \in \mathbb{N} \mid v^n_k = 0 \}. \]

(iv) Update according to

\[ u^{n+1}_I = 0 \]
\[ u^{n+1}_A = (\gamma K^* K|_{A \rightarrow A})^{-1}(v^n_A - u^n_A + \gamma (K^* K|_{A \rightarrow A} u^n_A + K^* K|_{I \rightarrow A} u^n_I)). \]

Here we denoted with \( K^* K|_{A \rightarrow A} \) and \( K^* K|_{I \rightarrow A} \) the operator \( K^* K \) restricted to the indicated sets. Note that step (iv) involves the inversion of the operator \( K^* K|_{A \rightarrow A} \) which is in general finite dimensional and hence, for injective \( K \), only an ill-conditioned problem and not ill-posed.

In [9] it is shown that the semismooth Newton method converges locally superlinear, i.e. if \( \| u^0 - u^* \| \) is small enough, it holds

\[ \frac{\| u^{n+1} - u^* \|}{\| u^n - u^* \|} \rightarrow 0. \]

4. How to check for convergence?

Many algorithms for the minimization of (1) come without a justified stopping rule. Two possibilities are described below:

(i) Motivated by (11) we can define the residual \( r^n = u^n - S_{w,1}(u^n - K^*(Ku^n - f)) \) and terminate the iteration while the residual is smaller that some tolerance \( \text{tol} \), i.e.

\[ \| r^n \| < \text{tol}. \]
(ii) In [1] a stopping rule based on an estimate for the distance to the minimal value of the functional is introduced. The functionals under consideration are of the form

$$\Psi(u) = \frac{\|Ku - f\|^2}{2} + \Phi(u) \quad (12)$$

with a proper, convex and lower semi-continuous $\Phi$ which also fulfills that $\Phi(u)/\|u\| \to \infty$ for $\|u\| \to \infty$. It is shown in [1] that the value

$$D(u) = \langle K^*(Ku - f), u \rangle + \Phi(u) + \Phi^*(-K^*(Ku - f)) \quad (13)$$

is an upper bound for the distance to the minimal value, i.e.

$$D(u) \geq \Psi(u) - \min_v \Psi(v).$$

Here we denoted with $\Phi^*$ the Fenchel dual of $\Phi$. Furthermore it is shown that $D(u^n) \to 0$ if $u^n$ is a sequence converging to a minimizer of $\Psi$. Hence, the value $D(u^n)$ can serve as a stopping criterion for any iterative procedure minimizing functionals of the form (12). Note that the condition on $\Phi$ is fulfilled for the functionals (8) for $1 \leq p \leq 2$.

5. Numerical experiments

In this section we illustrate the performance of the iterated soft thresholding from [4, 7, 17] and the iterated hard thresholding from [1]. Moreover we used a combined approach where iterated soft and hard thresholding are applied alternatingly. We also show an experiment on the semismooth Newton method.

5.1. Performance for different regularization parameters

In this experiment we analyzed the performance of the different algorithms for a fixed problem but for different regularization parameters. The problem under consideration is inverse integration (or differentiation [10, 16, 1]), i.e. the operator $K : L^2([0,1]) \to L^2([0,1])$ given by

$$Ku(t) = \int_0^t u(s)ds, \quad t \in [0,1].$$

The data $f$ is given as $(f(t_k))_{k=1,...,N}$ with $t_k = \frac{1}{N} k$. We discretized the operator $K$ by the matrix

$$K = \frac{1}{N} \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \ldots & \ldots & 1 \end{pmatrix}, \quad K : \mathbb{R}^N \to \mathbb{R}^N.$$

The minimization problem reads

$$\min_{u \in \mathbb{R}^N} \frac{1}{2} \sum_{i=1}^N ((Ku)_i - f^\delta)^2 + \sum_{k=1}^N w_k |u_k|.$$

The true solution $u^0$ is given by small plateaus and hence the data $f^\delta = Ku^0 + \delta$ is a noisy function with steep linear ramps. Figure 1 shows our sample data. The results of the $\ell^1$ minimization after 1000 iterations of the soft, hard and combined thresholding is shown in Figure 2 while Figure 3 and Table 1 report the obtained functional value and the residual norm.
Figure 1. The data which is used for the inverse integration experiment. Left: the true solution $u$ with $N = 500$, middle: the noisy data $f$ with 5% noise.

Figure 2. The results of 1000 iterations with regularization parameter $w = 0.0001$, left: iterated soft thresholding, middle: iterated hard thresholding, right: combination.

Figure 3. The evolution of the functional value $\Psi(u^n)$ during 1000 iterations with regularization parameter $w = 0.0001$, left: iterated soft thresholding, middle: iterated hard thresholding, right: combination.

| $w_k$   | $\Psi(u^n)$ soft | $\Psi(u^n)$ hard | $\Psi(u^n)$ combined | $\|r^n\|$ soft | $\|r^n\|$ hard | $\|r^n\|$ combined |
|---------|------------------|------------------|-----------------------|----------------|----------------|-------------------|
| 0.0100  | 7.38e+00         | 6.90e+00         | 6.89e+00              | 1.92e-02       | 1.14e-01       | 1.14e-01          |
| 0.0010  | 2.38e+00         | 1.15e+00         | 1.25e+00              | 3.15e-02       | 2.31e-02       | 2.31e-02          |
| 0.0001  | 1.61e+00         | 1.56e-01         | 1.70e-01              | 3.28e-02       | 2.51e-03       | 2.51e-03          |

Table 1. Comparison of the convergence of the iterated soft and hard shrinkage and their combination for different regularization parameters. The problem under consideration is inverse integration. For all cases we applied $n = 1000$ iterations.
Note that for a larger regularization parameter the iterated soft thresholding performs well concerning the residual norm, while for smaller values it gets worse. The combined approach is always good concerning the residual norm while for the smallest regularization parameter the iterated hard shrinkage gives the smallest residual norm. Another observation is that the functional value gives different results. Here the iterated hard thresholding is always good and the iterated soft thresholding performs comparably bad.

5.2. Scaling of the algorithms

In this experiment we assess how the computational cost and the number of iterations grow with the size of the problem. Again we considered the inverse integration problem from Subsection 5.1 with problem sizes from 32 to 512. We stopped the different algorithms when the residual norm $\|r^n\|$ falls below the value $5 \cdot 10^{-3}$. Table 2 and Figure 4 report the results. The combined approach is always better than the soft and the hard thresholding, both for CPU time and the number of iterations.

We emphasize that the stopping criterion was the residual norm. It turned out in our examples that the iterated hard thresholding reduces the functional value much faster than the iterated soft thresholding, see Figure 5 and Table 3 for an illustration of this effect. Figure 5 also reports the evolution of the estimator $D$ from (13). Note that this value decays much faster for the soft thresholding and the combined approach than for the the hard thresholding. We can conclude that neither the residual norm nor the estimator $D$ estimates the decay of the functional value properly: a small functional value may lead to a large residual norm and vice versa.

### Table 2. Comparison of the CPU time in seconds and the number of iterations to reach a given residual norm. The problem under consideration is inverse integration with 5\% noise and a regularization parameter $w_k = 5 \cdot 10^{-3}$.

| $N$ | soft  | hard  | combined | soft  | hard  | combined |
|-----|-------|-------|----------|-------|-------|----------|
| 32  | 5.32e-01 | 1.80e+00 | 4.35e-01 | 422  | 5095  | 176      |
| 64  | 1.58e+00 | 6.39e+02 | 1.22e+00 | 4556 | 2060216 | 2932     |
| 128 | 2.44e+00 | 6.38e+02 | 1.57e+00 | 6233 | 1539026 | 3022     |
| 256 | 3.72e+00 | 1.72e+03 | 3.28e+00 | 4290 | 1687510 | 3208     |
| 512 | 1.60e+01 | 6.28e+03 | 1.17e+01 | 7841 | 2167682 | 4624     |

### Figure 4. Same data as in Table 2. Left: CPU times in second, right: number of iterations.
Figure 5. Left: The evolution of the functional value $\Psi(u^n)$ for the different methods. Middle: The evolution of the residual norm $||r^n||$ for the different methods. Right: The evolution of the value $D(u^n)$ for the different methods. The problem under consideration is the inverse integration with $N = 512$, 5% noise and $w_k = 10^{-2}$. Note that the residual norm decreases faster for soft than for hard thresholding, but the functional value behaves the other way round. The combined approach is always best.

| $\Psi$  | soft | hard | combined |
|---------|------|------|----------|
| 6.000   | 276  | 20   | 28       |
| 5.500   | 493  | 69   | 41       |
| 5.000   | 990  | 126  | 136      |
| 4.500   | 2646 | 333  | 484      |
| 4.190   | 7694 | 854  | 1220     |

Table 3. Number of iterations to reach a given functional value $\Psi$ for the different methods. The problem is the same than in the last line of Table 2. The minimal value of $\Psi$ in this problem is 3.94.

5.3. Deblurring in a Haar basis

As an example for the performance of the semismooth Newton method from Section 3 we consider a blurring operator $A : L^2([0,1]) \to L^2([0,1])$ given by $Au = k * u$ with the kernel $k(x) = c (1 + x^2/\lambda^2)^{-1}$ with $\lambda = 0.01$. We choose $c$ such that $\int k \, dx = 1$ and consider $u$ to be extended periodically to $\mathbb{R}$ in order to evaluate the convolution integral.

In this example we work with a synthesis operator $B : \ell^2 \to L^2([0,1])$ mapping coefficients $(c_k)$ to a function $u = \sum_k c_k \psi_k$ where the $(\psi_k)$ form the orthonormal Haar wavelet basis [13]. Hence, the operator under consideration $K = AB$ is a blurring after a Haar wavelet synthesis.

We start with a given function $u$ which is piecewise constant. The data $f$ is computed as $f = Au + \text{noise}$ such that we have 25% relative error, i.e. $||f - Au||/||f|| = 0.25$. The Haar coefficients of $u$ have been reconstructed by minimizing (1). As a comparison we also computed the classical Tikhonov regularization, which amounts to the minimization of

$$\frac{1}{2} ||Kc - f||^2 + \sum_{k=1}^{\infty} w_k |c_k|^2.$$  

Figure 6 and Table 4 show the results of both $\ell^1$ and the above $\ell^2$ regularization where we discretized the problem to 1024 Haar wavelets. The parameters $w_k$ have been tuned by hand to produce optimal results. Since the original data is quite sparse in the Haar wavelet basis, the $\ell^1$-reconstruction leads to much better results. It also turned out that the algorithm is robust with respect to different initial values $u^0$. We tested several initial values (starting at zero, at $K^*f$ or at a random position) and the observed convergence behavior was very similar in all cases.
The SSN method converged in six iterations and in 0.3 seconds (for comparison: soft thresholding converged in 329 iterations in 2.8 seconds, hard thresholding did not converge within a reasonable time and the combined approach needed 412 iterations and 3.3 seconds).

Figure 6. The results of $\ell^1$ and $\ell^2$ (classical Tikhonov) regularization of deblurring in a Haar basis. Upper left: the true solution $u$, upper right: the given data $f$, lower left: the reconstruction by $\ell^1$ minimization with $w_k = 0.12$ and $\gamma = 5 \cdot 10^6$, lower right: the reconstruction by $\ell^2$ minimization with $w_k = 0.05$.

| $n$ | $\Psi(u^n)$ | $\|r^n\|$ |
|-----|-------------|----------|
| 1   | 3.3920e+01  | 2.9676e+06 |
| 2   | 1.3905e+02  | 3.2499e+04 |
| 3   | 1.3326e+01  | 6.2647e+05 |
| 4   | 7.9347e+00  | 1.7517e+04 |
| 5   | 6.0006e+00  | 8.0510e-02 |
| 6   | 5.9823e+00  | 5.7542e-09 |

Table 4. Illustration of the performance of the SSN method for deblurring in a Haar basis. The second column shows the decay of the function value $\Psi$ while the third column shows the norm of the residual. The data is the same than in Figure 6.

6. Conclusion
We have seen that the well known iterated soft thresholding for minimization problems with sparsity constraint can be seen as either a generalized projection gradient method or a generalized
conditional gradient method. Furthermore we summarized results on iterated hard thresholding as a generalized conditional gradient method and a semismooth Newton method for the same problem. While for both soft and hard thresholding one only needs to apply the operator and its adjoint once for each iteration in the semismooth Newton method a finite (probably small) linear system of equations has to be solved.

In our numerical experiments we also investigated an approach which combines the soft and the hard thresholding by simply using them alternatingly. It turned out that this combination almost always improves the convergence when measured in different terms like the residual norm, the functional value on the estimator $D$ from (13). The semismooth Newton method converges fast but no global convergence result is available at the moment.

References

[1] Kristian Bredies and Dirk A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. To appear in *SIAM Journal on Scientific Computing*, 2007.
[2] Kristian Bredies and Dirk A. Lorenz. Iterated soft-thresholding converges linearly. Preprint, 2007.
[3] Kristian Bredies, Dirk A. Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. To appear in *Computational Optimization and Applications*, 2005.
[4] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications in Pure and Applied Mathematics*, 57(11):1413–1457, 2004.
[5] Ingrid Daubechies, Massimo Fornasier, and Ignace Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. Preprint arXiv:0706.4297v1, 2007.
[6] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.
[7] Mário A. T. Figueiredo and Robert D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.
[8] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Applications to compressed sensing and other inverse problems. Preprint, 2007.
[9] Roland Griess and Dirk A. Lorenz. A semismooth Newton method for Tikhonov functionals with sparsity constraints. Preprint, 2007.
[10] Martin Hanke and Otmar Scherzer. Inverse problems light: Numerical differentiation. *The American Mathematical Monthly*, 108(6):512–521, 2001.
[11] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003.
[12] Dirk A. Lorenz. Non-convex variational denoising of images: Interpolation between hard and soft wavelet shrinkage. *Current Development in Theory and Application of Wavelets*, 1(1):31–56, 2007.
[13] Alfred Karl Louis, Peter Maass, and Andreas Rieder. *Wavelets: Theory and Application*. Wiley, Chichester, 1997.
[14] Ronny Ramlau and Gerd Teschke. A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. *Numerische Mathematik*, 104(2):177–203, 2006.
[15] R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*, Springer, 1998.
[16] Frank Schöpfer, Alfred K. Louis, and Thomas Schuster. Nonlinear iterative methods for linear ill-posed problems in Banach spaces. *Inverse Problems*, 22:311–329, 2006.
[17] Jean-Luc Starck, Mai K. Nguyen, and Fionn Murtagh. Wavelets and curvelets for image deconvolution: a combined approach. *Signal Processing*, 83:2279–2283, 2003.
[18] Gerd Teschke. Multi-frame representations in linear inverse problems with mixed multi-constraints. *Applied and Computational Harmonic Analysis*, 22(1):43–60, 2007.