Radon Series Comp. Appl. Math **XX**, 1–62

Sparse Recovery in Inverse Problems

Ronny Ramlau and Gerd Teschke

Abstract. Within this chapter we review some recent results on sparse recovery algorithms in the context of inverse and ill-posed problems. The review centers in particular on those inverse problems in which we can assume that the solution has a sparse series expansion with respect to a preassigned basis or frame. The presented approaches to approximate solutions of inverse problems are limited to iterative strategies that essentially rely on the minimization of Tikhonov-like variational problems, where the sparsity constraint is integrated through ℓ_p norms. In addition to algorithmic and computational aspects, we also discuss in greater detail regularization properties that are required for cases in which the operator is ill-posed and no exact data are given. Such scenarios reflect realistic situations and manifest therefore its great suitability for "real-world" applications.

The material reviewed here originates from very recent work of the authors of this chapter. The focus is on iterated soft-shrinkage and projected steepest descent for nonlinear inverse problems and nonlinear approximation technologies for linear inverse problems. For all proposed algorithms we provide regularization results, convergence rates (if possible), and numerical examples.

Keywords. inverse and ill-posed problems, regularization theory, convergence rates, sparse recovery, iterated soft-shrinkage, accelerated steepest descent, nonlinear approximation.

AMS classification. 41A46, 45Q05, 47A52, 49M99, 52A99, 65B99, 65K10.

The research of G.T. was partially supported by Deutsche Forschungsgemeinschaft Grants TE 354/3-1, TE 354/4-1, TE 354/5-1, TE 354/8-1.

1 Introduction to Classical Inverse Problems

In many applications in the natural sciences, medicine or imaging one has to determine the cause x of a measured effect y. A classical example is Computerized Tomography (CT), a medical application, where a patient is screened using x - rays. The observed damping of the rays is then used to reconstruct the density distribution of the body. In order to achieve such a reconstruction, the measured data and the searched for quantity have to be linked by a mathematical model, which we will denote by F (or A, if the model is linear). In an abstract setting, the determination of of the cause x can be stated as follows: Solve an operator equation

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

$F : X \to Y$, where X, Y are Banach (Hilbert) spaces. For the CT problem, the operator describing the connection between the measurements and the density distribution (in 2 dimensions) is given by the Radon transform,

$$y(s,\omega) = (Ax)(s,\omega) = \int_{\mathbb{R}} x(s\omega + t\omega^{\perp})\, dt \ , \ s \in \mathbb{R}, \ \omega \in \mathcal{S}^1 .$$

As in practice the observed data stems from measurements, one never has the exact data y available, but rather a noisy variant y^δ. In the following we might assume that at least a bound δ for the noise is available (e.g. if the accuracy of the measurement device is known):

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

In connection with Inverse Problems, the following questions arise:

(i) Does there exist a solution of equation (1.1) for given exact y?

(ii) Is the solution unique?

(iii) If the solution is determined from noisy data, how accurate is it?

(iv) How to solve (1.1)?

1.1 Preliminaries

In order to give a first idea on the problems that may be encountered for ill-posed problems, we will now consider a linear operator equation in finite dimensions. Assume $A \in \mathbb{R}^{n,n}$, and we want to solve the matrix equation $Ax = y$ from noisy data y^δ. If we assume that A is invertible on $R(A)$ and also $y^\delta \in R(A)$ (which is already a severe restriction), then we can define

$$\begin{aligned} x^\dagger &:= A^{-1}y \\ x^\delta &:= A^{-1}y^\delta \ . \end{aligned}$$

With $x - x^\delta = A^{-1}(y - y^\delta)$ the distance between $x, x^\dagger$ can be estimated as:

$$\begin{aligned} \|x - x^\delta\| &\leq \|A^{-1}\| \, \|y - y^\delta\| \\ &\leq \|A^{-1}\| \, \delta \, . \end{aligned} \tag{1.2}$$

If we additionally assume that A is symmetric and positive definite with $\|A\| \leq 1$, then A has an eigensystem (λ_i, x_i) with eigenvalues $0 < \lambda_i \leq 1$ and associated eigenvectors x_i. Moreover we have

$$\|A^{-1}\| = \frac{1}{\lambda_{min}} \Rightarrow \|x - x^\dagger\| \leq \frac{\delta}{\lambda_{min}}$$

Therefore, the reconstruction quality is of the same order $O(\delta)$ as the data error, magnified only by the norm of the inverse operator. However, it turns out that $O(\delta)$ estimates are only possible in a finite dimensional setting: Indeed, if we define the operator

$$Ax = \sum_{i=1}^{\infty} \lambda_i \langle x, x_i \rangle x_i$$

with orthonormal basis x_i and $\lambda_i \to 0$, then it is easily to see that the right hand side of estimate (1.2) explodes. In fact, for general inverse problems with $dim\ R(A)$ the best possible convergence rate is given by

$$\|x - x^\delta\| = O(\delta^s) \, , \quad s < 1. \tag{1.3}$$

The above considerations were based on the assumption $y^\delta \in R(A)$. As we will see in the following example, this is a severe restriction that will not hold in practice:
Let us consider the integral equation

$$y(s) = \int_0^s x(t)dt \qquad 0 \leq s \leq 1.$$

It follows immediately $y \in C^1[0, 1]$ and

$$x(s) = y'(s), \qquad y(0) = 0 \, .$$

For noisy measurements this condition will not hold, as the noise will not only alter the initial value but also the smoothness of y^δ, as the data noise is usually not differentiable. The same also holds for Computerized Tomography: It can be shown [Natterer] that the exact CT data belongs to the Sobolev space $H^{1/2}(\mathbb{R} \times \mathcal{S}^1)$, but for the noisy data we only have $y^\delta \in L_2$.

Now let us define *well-posed* and *ill-posed problems*.

Definition 1.1. Let $A : X \to Y$ linear operator and X, Y topological spaces. Then the problem (A, X, Y) is well-posed if condition (i)-(iii) are fulfilled at the same time,

(i) $Ax = y$ has a solution for each $y \in Y$

(ii) the solution is unique

(iii) the solution depends continuously on the data, i.e.

$$y_n \to y,\ y_n = F(x_n), \Longrightarrow x_n \to x \text{ and } F(x) = y\ .$$

If one of the conditions is violated, then the problem is ill posed.

Roughly speaking, well-posed problems allow for an error estimate as in (1.2), whereas the best possible rate for ill posed problems is as in (1.3).

Let us denote by $L(X, Y)$ the set of all linear and continuous operators $A : X \to Y$. An important class of operators that lead to ill-posed problems are *compact* operators.

Definition 1.2. An operator $A \in L(X, Y)$ is compact, if it maps bounded sets to relative compact sets, i.e., $\overline{R(B)}$ is compact for bounded sets B. Or equivalently, for any bounded sequence $\{x_n\}_{n\in\mathbb{N}}$, the sequence $y_n = Ax_n$ has a convergent subsequence.

Integral operators are an important class of examples for compact operators.

Definition 1.3. Let $G \in \mathbb{R}^n$ be a bounded set and $k : G \times G \to G$. We define the integral operator K by

$$(Kx)(s) = \int_G k(s,t)x(t)dt\ .$$

Proposition 1.4. *Let $k \in C(G, G)$ and K be an integral operator considered between any of the spaces $L_2(G)$ and $C(G)$. Then K is compact. If $k \in L_2(G, G)$, then the integral operator $K : L_2(G) \to L_2(G)$ is compact.*

Another example for compact operators are Sobolev embedding operators. For bounded G and a real number $s > 0$, let us consider the map

$$i_s : H^s(G) \to L_2(G)\ , \quad \text{which is defined by } \ i_s x = x\ .$$

Here H^s denotes the standard Sobolev space. Then we have

Proposition 1.5. *The Sobolev embedding operator i_s is compact.*

Proposition 1.6. *Compact operators with* $\dim R(K) = \infty$ *are not continuously invertible, i.e. they are ill-posed.*

Now let us assume that a given operator $A : H^s \to H^{s+t}$, $s \geq 0, t > 0$, is continuously invertible. As pointed out above, the measured data will not belong to H^{s+t} but rather to L_2. Therefore, we have to consider the operator equation between H^s and L_2, i.e. the equation $y = i_s(Ax)$. As a combination of a continuous and a compact operator, $i_s(A)$ is also compact and therefore not continuously invertible - regardless of the invertibility of A.

A key ingredient for the stable inversion of compact operators is the spectral decomposition:

Proposition 1.7. *Let $K : X \to X$, X a Hilbert space and assume that K is compact and self-adjoint (i.e, $\langle Kx, y \rangle = \langle x, Ky \rangle \; \forall x, y \in X$). By (λ_j, u_j) denote the set of eigenvalues λ_j and associated eigenvectors u_j with $Ku_j = \lambda_j u_j$. Then $\lambda_j \to 0$ (if $\dim R(K) = \infty$) and the functions u_j form an orthonormal basis of $\overline{R(K)}$ with*

$$Kx = \sum_{i=1}^{\infty} \lambda_i \langle x, u_i \rangle u_i \; .$$

The eigenvalue decomposition can be generalized to compact operators that are not self-adjoint. Let $K : X \to Y$ be given. The adjoint operator $K^* : Y \to X$ is formally defined by the equation

$$\langle Kx, y \rangle = \langle x, K^* y \rangle \quad \forall x, y \; .$$

We can then define the operator $K^*K : X \to X$ and find

$$\begin{aligned} \langle K^*Kx, y \rangle &= \langle Kx, Ky \rangle = \langle x, K^*Ky \rangle \; , \\ \langle K^*Kx, x \rangle &= \langle Kx, Kx \rangle = \|Kx\|^2 \; , \end{aligned}$$

i.e., K^*K is selfadjoint and positive semi-definite, which also guarantees that all eigenvalues λ_i of K^*K are nonnegative. Therefore we have

$$K^*Kx = \sum_i \lambda_i \langle x, u_i \rangle u_i \; .$$

Defining

$$\begin{aligned} \sigma_i &= +\sqrt{\lambda_i} \\ Ku_i &= \sigma_i v_i \; , \end{aligned}$$

we find that the functions v_i also form an orthonormal system for X:

$$\begin{aligned} \langle v_i, v_j \rangle &= \frac{1}{\sigma_i \sigma_j} \langle Ku_i, Ku_j \rangle \\ &= \frac{1}{\sigma_i \sigma_j} \langle K^*Ku_i, u_j \rangle \\ &= \frac{\sigma_i}{\sigma_j} \langle u_i, u_j \rangle = \delta_{ij} = \begin{cases} 1 \; i = j \\ 0 \; i \neq j \end{cases} \; , \end{aligned}$$

and get

$$\begin{aligned} Kx &= K(\sum_i \langle\, x, u_i \,\rangle u_i) = \sum_i \langle\, x, u_i \,\rangle K u_i = \sum_i \sigma_i \langle\, x, u_i \,\rangle v_i \;, \\ K^* y &= \sum_i \sigma_i \langle\, y, v_i \,\rangle u_i \;. \end{aligned}$$

The above decomposition of K is called the *singular value decomposition* and $\{\sigma_i, x_i, y_i\}$ is the singular system of K. The *generalized inverse* of K is defined as follows:

Definition 1.8. The generalized inverse $K^\dagger$ of K is defined as

$$\begin{aligned} D(K^\dagger) &= R(K) \oplus R(K)^\perp \\ K^\dagger y &:= x^\dagger \\ x^\dagger &= \arg\min_x \|y - Kx\| \;. \end{aligned}$$

If the minimizer $x^\dagger$ of the functional is not unique then the one with minimal norm is taken.

Proposition 1.9. *The generalized solution $x^\dagger$ has the following properties*

(i) *$x^\dagger$ is the unique solution of $K^*Kx = K^*y$,*

(ii) *$Kx^\dagger = P_{R(K)}y$, where $P_{R(K)}$ denotes the orthogonal projection on the range of K,*

(iii) *$x^\dagger$ can be decomposed w.r.t. the singular system as*

$$x^\dagger = \sum_i \frac{1}{\sigma_i} \langle y, v_i \rangle u_i \;, \tag{1.4}$$

(iv) *the generalized inverse is continuous if and only if $R(K)$ is closed.*

A direct consequence of the above given representation of $x^\dagger$ is the so-called Picard condition:

$$y \in R(K) \Leftrightarrow \sum_i \frac{|\langle\, y, v_i \,\rangle|^2}{\sigma_i^2} < \infty \;.$$

The condition states that the moments of the right hand side y (w.r.t. to the function system $\{v_i\}$) have to tend to zero fast enough in order to compensate the growth of $1/\sigma_i$.

What happens if we apply noisy data to formula (1.4)? Assume $y \in R(K)$, $y = Kx^\dagger$, and $y_l^\delta = y + \delta v_l$. Then for all l

$$\|y - y_l^\delta\| \le \delta \;,$$

but with

$$x^\delta = \sum_i \frac{1}{\sigma_i} \langle y_l^\delta, v_i \rangle u_i$$

we obtain

$$\|x - x^\delta\|^2 = \sum_i \frac{\delta^2}{\sigma_i^2} |\langle v_l, v_i \rangle|^2 = \frac{\delta^2}{\sigma_l^2} \to \infty \text{ as } l \to \infty ,$$

which shows that the reconstruction error can be arbitrarily large even if the noisy data is close to the true data.

1.2 Regularization Theory

In order to get a reasonable reconstruction, we have to introduce different methods that ensure a good and stable reconstructions. These methods are often defined via functions of operators.

Definition 1.10. Let $f : \mathbb{R}^\dagger \to \mathbb{R}$. For compact operators, we define

$$f(K)x := \sum_i f(\sigma_i) \langle x, u_i \rangle v_i .$$

Of course, this definition is only well-defined for functions f for which the sum converges. We can now define *regularization methods*.

Definition 1.11. A regularization of an operator $K^\dagger$ is a family of operators $\{R_\alpha\}_{\alpha>0}$,

$$R_\alpha : Y \to X$$

with the following properties: there exists a map $\alpha = \alpha(\delta, y^\delta)$ such that for all $y \in D(K^\dagger)$ and all $y^\delta \in Y$ with $\|y - y^\delta\| \le \delta$,

$$\lim_{\delta \to 0} R_{\alpha(\delta, y^\delta)} y^\delta = x^\dagger$$

and

$$\lim_{\delta \to 0} \alpha(\delta) = 0 .$$

The parameter α is called regularization parameter.

In the classical setting, regularizing operators R_α are defined via filter functions F_α:

$$R_\alpha y^\delta := \sum_{i \in \mathbb{N}} \sigma_i^{-1} F_\alpha(\sigma_i) \langle y^\delta, v_i \rangle u_i .$$

The requirements of Definition 1.11 have some immediate consequences on the admissible filter functions. In particular, $D(R_\alpha) = Y$ enforces $|\sigma_i^{-1} F_\alpha(\sigma_i)| \le C$ for all i, and the pointwise convergence of R_α to $K^\dagger$ requires $\lim_{\alpha \to 0} F_\alpha(t) = 1$. Well-known regularization methods are:

(i) Truncated singular value decomposition:

$$R_\alpha y^\delta := \sum_i^N \sigma_i^{-1} \langle y^\delta, v_i \rangle u_i$$

In this case, the filter function is given by

$$F_\alpha(\sigma) := \begin{cases} 1 & \sigma \geq \alpha \\ 0 & \sigma < \alpha \end{cases} .$$

(ii) Truncated Landweber iteration: For $\beta \in (0, \frac{2}{\|K\|^2})$ and $m \in \mathbb{N}$, set

$$F_{1/m}(\lambda) = 1 - (1 - \beta\lambda^2)^m .$$

Here, the regularization parameter $\alpha = 1/n$ only admits discrete values.

(iii) Tikhonov regularization: Here, the filter function is given by

$$F_\alpha(\sigma) = \frac{\sigma^2}{\sigma^2 + \alpha} .$$

The regularized solutions of Landweber's and Tikhonov's method can also be characterized as follows:

Proposition 1.12. *The regularized solution due to Landweber, $x^\delta_{1/m}$, is also given by the m-th iterate of the Landweber iteration given by*

$$x_{n+1} = x_n + K^*(y^\delta - Kx_n) , \quad \text{with } x_0 = 0 \quad .$$

The regularization parameter is the stopping index of the iteration.

Proposition 1.13. *The regularized solution due to Tikhonov,*

$$x^\delta_\alpha := \sum_i \frac{\sigma_i^2}{\sigma_i^2 + \alpha} \cdot \sigma_i^{-1} \langle y^\delta, v_i \rangle u_i ,$$

is also the unique minimizer of the Tikhonov functional

$$J_\alpha(x) = \|y^\delta - Kx\|^2 + \alpha\|x\|^2 , \tag{1.5}$$

which is minimized by the unique solution of the equation

$$(K^*K + \alpha I)x = K^* y^\delta .$$

Tikhonov's variational formulation (1.5) is important as it allows generalizations towards nonlinear operators as well as to sparse reconstructions. As mentioned above, regularization methods also require proper parameter choice rules.

Proposition 1.14. *The Tikhonov regularization combined with one of the parameter choice rules*

a) $\alpha(\delta) \to 0$ *and* $\frac{\delta^2}{\alpha(\delta)} \to 0$

b) $\alpha_*(\delta, y^\delta)$ *s.t.* $\|y^\delta - Kx^\delta_{\alpha_*}\| = \tau\delta$ *for fixed* $\tau > 1$ *(discrepancy principle)*

is a regularization method.

Proposition 1.15. *Let* $\tau > 1$. *If the Landweber iteration is stopped after* m_* *iterations, where* m_* *is the first index with*

$$\|y^\delta - Kf_{m_*}\| \le \tau\delta < \|y^\delta - Kf_{m*-1}\| \quad (discrepancy\ principle)\ ,$$

then the iteration is a regularization method with $R_{\frac{1}{m_*}} y^\delta = f_m*$.

The last two propositions show that the regularized solutions for Tikhonov's or Landweber's method converge towards the true solution provided a proper parameter choice rule was applied. However, no result on the speed of convergence is provided. Due to Bakhushinsky one rather has

Proposition 1.16. *Let* $x^\delta_\alpha = R_\alpha y^\delta, R_\alpha$ *be a regularization method. Then the convergence of* $x^\delta_\alpha \to x^\dagger$ *can be arbitrary slow.*

To overcome this drawback, we have to assume a certain regularity of the solution. Indeed, convergence rates can be achieved provided the solution fulfills a so-called source-conditions. Here we limit ourselves to the Hölder-type source conditions,

$$x^\dagger = (K^*K)^{\frac{\nu}{2}} w, \text{ i.e. } x^\dagger \in R(K^*K)^{\frac{\nu}{2}} \subseteq D(K) = X, \nu > 0\ .$$

Definition 1.17. A regularization method is called order optimal if for a given parameter choice rule the estimate

$$\|x^\dagger - x^\delta_{\alpha(\delta)}\| = 0(\delta^{\frac{\nu}{\nu+1}})$$

holds for all $x^\dagger = (K^*K)^{\frac{\nu}{2}} w$ and $\|y^\delta - y\| \le \delta$.

It turns out that for $x^\dagger = (K^*K)^{\frac{\nu}{2}} w$ this is actual the best possible convergence rate, no method can do better. Also, we have $\delta^{\frac{\nu}{\nu+1}} > \delta$ for $\delta < 1$, i.e., we always loose some information in the reconstruction procedure.

Proposition 1.18. *Tikhonov regularization and Landweber iteration together with the discrepancy principle are order optimal.*