Image Reconstruction Through Regularization by Envelope Guided Conjugate Gradients

Linda Kaufman and Arnold Neumaier

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974-0636
U.S.A.

email: lck@research.att.com, neum@cma.univie.ac.at

Abstract

In this paper we propose a new way to iteratively solve large scale ill-posed problems, and in particular the image reconstruction problem from noisy images or noisy data linearly related to the pixel intensities. This is done by exploiting the relation between Tikhonov regularization and multiobjective optimization to obtain iteratively approximations to the Tikhonov L-curve and its corner. Monitoring the change of the approximate L-curves allows us to adjust the regularization parameter adaptively during a preconditioned conjugate gradient iteration, so that the desired image can be reconstructed with a low number of iterations. Nonnegativity constraints are taken into account automatically. We present test results on image reconstruction in positron emission tomography (PET).

Keywords: Tikhonov regularization, multiobjective optimization, ill-posed, L-curve, envelope, preconditioned conjugate gradients, image reconstruction, positron emission tomography (PET)

1991 MSC Classification: primary 65F10, secondary 65R30, 68U10, 90C29, 92C55

1 Introduction

Many problems in applied mathematics lead to models of the form

$$F(x) = y + \epsilon,$$

where x is an unknown vector of parameters, often restricted to a subset $\Omega \subset \mathbb{R}^n$ (e.g., by nonnegativity constraints $x \geq 0$) to be determined from a data vector y. The term ϵ is a small noise term, often of unknown magnitude, which accounts for systematic model errors and random stochastic errors.

Frequently, the solution of $F(x) = y$ depends very sensitively on y, or is not even uniquely determined by it. This is particularly noticeable in discretized versions of ill-posed problems which often arise from so-called 'inverse' problems for which standard existence and uniqueness arguments break down. To get reliable and robust estimates for x from the measured data y, one needs to exploit the availability of qualitative information about x, which usually is given in the form of a vague statement that some measure of smoothness of x is not large.

Sources of ill-posed problems include image smoothing, deconvolution (Wiener filtering), shape from shading, computer-assisted tomography (CAT, PET), indirect measurement, nondestructive testing, inverse scattering, seismic analysis, parameter identification in dynamical systems, analytic continuation, inverse Laplace transform, relaxation spectra, and partial differential equations of mixed type (e.g., multiphase flow) or with nonstandard boundary conditions (e.g., backward heat equation). A recent survey by Engl [2] gives details and references for a number of such problems, and an overview of the analytical results available. Hanke & Hansen [6] survey the numerical analysis of methods for approximating the solution of ill-posed problems.

In the present paper we introduce a new aspect to the analysis of ill-posed problems by looking at it from the point of view of multiobjective optimization. Indeed, the standard regularization techniques can be viewed as compromises designed to make small both a measure $r(x)$ for the lack of fit and a measure $q(x)$ for the lack of smoothness. This new point of view results in new numerical techniques for the robust iterative solution of high-dimensional ill-posed problems, which arise especially in problems where two- or three-dimensional functions or images need to be determined.

Our research was motivated by (and tested with) the image reconstruction problem in positron emission tomography, but it is not limited to that problem.

2 Regularization techniques in image reconstruction

Image reconstruction is the problem of finding a piecewise smooth image x from data derived from this image (or rather the corresponding real life original). The data may be a noisy version of the image, an incomplete version of such a noisy image, an image degraded by an approximately known blurring operator, or projection data of various sorts.

For example, in positron emission tomography (PET) the patient is given a tagged substance (such as glucose for brain study) that emits positrons. Each positron annihilates with an electron and emits two photons in opposite directions. The patient is surrounded by a ring of detectors, which are wired so that whenever any pair of detectors senses a photon within a very small time interval, the size of which is system-dependent, the count for that pair is incremented. In a matter of minutes several million photon pairs may be detected. The reconstruction problem in PET is to determine a memory map of the annihilations, and hence a map of the blood flow, given the data gathered by the ring of detectors.

One of the problems of practical interest is to find methods that allow the image reconstruction from the fewest possible number of photons, since this allows more flexible use of the method and the application to a wider range of problems. Unfortunately, the noise level increases with a decreasing number of recorded photon pairs, and many reconstruction methods deteriorate quickly with increasing noise.

Assume one has imposed a grid of boxes on the affected organ and tries to compute the unknown number x_B of annihilations in the box B. Let b_d represent the number of photon pairs detected in tube d and let the matrix entry $A_{d,B}$ represent the probability that an emission in box B is detected in tube d. Then the desired density x satisfies the approximate relation

$$Ax \approx b.$$

(A careful discussion of the details of the linear model used and the underlying assumptions can be found in VARDI et al. [26]. The comments to that paper also contain interesting remarks on corrections needed in actual practice.)

The probability matrix approach of SHEPP & VARDI [20] to the reconstruction problem is plagued by a snowy image problem. This is particularly noticed when there are relatively few photon pairs detected – say under 5 million. The snow is attributed to the fact that the algorithm is doing a good fit to the noise in the data as well as to the signal in the data, and since the problem is ill-conditioned, fitting the noise has a considerable influence on the quality of the image.

Various suggestions have been made to remedy this problem, and the resulting techniques are usually referred to as *regularization* methods. Since initially the EM algorithm does a good job of fitting the signal to the data, some people suggest stopping the algorithm as soon as the rate that function being optimized decreases significantly. The maximum likelihood function values traced out by the EM algorithm tend to resemble a bent leg and one would stop as soon as one is over the leg's knee. With various accelerating schemes discussed in KAUFMAN [13], the rise is faster and it is simpler to determine when there is a change in the slope of the function. This is particularly noticeable if one substitutes a least squares merit function for the maximum likelihood merit function and uses a preconditioned conjugate gradient approach, as in KAUFMAN [14].

The main problem with the technique of *regularization by truncating the iteration* is the difficulty to automate the decision of when to stop the iteration. Missing the 'right' stopping point may result in a significant degradation of the resulting image quality. The state of the art of finding and analyzing stopping rules is surveyed in Section 6 of HANKE & HANSEN [6]. Moreover, as we shall see, even truncating at the optimal iteration which gives the minimal mean squared error (for simulated problems where one can compute this) may yield a reconstructed image of poorer quality than necessary.

A more versatile regularization technique, used also for other ill-posed inverse problems, entails adding a smoothing penalty term to the objective function. In the penalty approach one adds to the squared Euclidean norm of the residual,

$$r(x) = ||Ax - b||_2^2, \tag{1}$$

a multiple of a penalty term $q(x)$ measuring the roughness of the image, and thus minimizes

$$f(x) = r(x) + \lambda q(x), \tag{2}$$

where λ is a scalar penalty parameter, usually referred to as the *regularization parameter*.

With this formulation one can also handle the problem of finding a smoothed version x of a noisy image b by taking A as the identity matrix, of completing an incomplete noisy image b with missing pixels to a full smoothed image x by taking for A a (0,1)-matrix with ones in the positions of the known pixels, and of deblurring an image B degraded by a known linear blurring operator, which defines A.

In the context of maximum likelihood methods, this technique is derived from a Markov random field approach and known as the *maximum a posteriori* or MAP procedure, and the penalty term is then referred to as the *Gibbs prior*, see, e.g., LALUSH & TSUI [15]. Statisticians usually refer to the special case where

$$q(x) = \|x\|_2^2, \tag{3}$$

as *ridge regression*. While (3) is the simplest possible choice of the penalty function $q(x)$, it does not reflect a priori knowledge about the image, such as the expectation that large parts of the image are likely to be fairly smooth. To take account of this, various suggestions have been given, including the (convex, quadratic) function

$$q(x) = \sum_j \left(\frac{1}{8} \sum_{i \sim j} x_i - x_j \right)^2, \tag{4}$$

where $i \sim j$ denotes the statement that i is adjacent to j (i.e., is one of the 8 neighbors of j on the grid), the (asymptotically constant) rational function

$$q(x) = \sum_j \sum_{i \sim j} \frac{(x_i - x_j)^2}{(x_i - x_j)^2 + \delta}, \tag{5}$$

suggested in GEMAN & MCCLURE [4], the (asymptotically sublinear) function of HEBERT & LEAHY [9]

$$q(x) = \sum_j \sum_{i \sim j} \log \left(1 + \frac{(x_i - x_j)^2}{\delta} \right), \tag{6}$$

the suggestion in GREEN [5] of the (convex, asymptotically linear) function

$$q(x) = \sum_j \sum_{i \sim j} \log\cosh\left(\frac{x_i - x_j}{\delta}\right) \tag{7}$$

and the (convex, asymptotically linear) multiquadric function

$$q(x) = \sum_j \sum_{i \sim j} \sqrt{(x_i - x_j)^2 + \delta}, \tag{8}$$

in VOGEL & OMAN [27], which all have an additional parameter δ. Note that when the image is represented as a two-dimensional surface, (4) is a cumulative measure of local curvature since it vanishes for planar surfaces, while (5) – (8) are cumulative measures of local steepness. The relative merits of the penalty functions (5) – (7) were discussed in LALUSH & TSUI [15].

One notices the similarities with robust regression (HUBER [12]). Indeed, one of the reasons for using nonquadratic penalty terms is the wish to allow a certain amount of nonsmoothness along edges of the image. Such edges contribute an unduly large amount to quadratic penalties, while other nonlinearities are more tolerant. Thus the contribution of the edges to the penalty term behave like the contributions of outliers to the regression terms, and in robust regression, the latter are similarly accounted for by nonlinear terms in place of the least squares terms. Motivated by this, BLAUER & LEVINE [1] and STEVENSON et al. [21], also looked at the penalty term

$$q(x) = \sum_j \sum_{i \sim j} H(x_i - x_j, \delta) \tag{9}$$

(and at more elaborate versions also incorporating local curvature information), where is H is the piecewise quadratic-linear function

$$H(d, \delta) = \begin{cases} d^2 & \text{if } |d| < \delta, \\ 2\delta|d| - \delta^2 & \text{otherwise} \end{cases} \tag{10}$$

introduced by HUBER [12] in the context of robust regression.

In passing we also note that there are methods that regularize with discontinuous penalty functions. GEMAN & GEMAN [3] require the use of expensive simulated annealing techniques for minimizing the resulting objective function, but their approach gives excellent results for piecewise constant images. Another approach, which does not quite fit into the above setting, is that of Terzopoulos [23], that adaptively restricts the sum in a quadratic penalty function to neighbors not separated by a discontinuity; the locations of the discontinuities are adapted by further penalty terms.

With any of the penalty functions discussed, a value of λ can be determined that gives a picture that is esthetically pleasing. However, as shown in Kaufman [14] and Lalush & Tsui [15], this value is strongly data dependent, and determining it by trial and error for each image is rather time consuming. The image processing literature contains little information on how to adjust the regularization parameter λ, and most studies don't explain why they chose the particular regularization parameter they were using. Exceptions are Reeves & Mersereau [18], who discuss selecting λ by a simplified generalized cross-validation approach, and Hebert & Leahy [10], who change λ adaptively during the iteration using a statistical feedback scheme based on the chi-square test.

The problem of choosing λ has been looked at more closely by the larger applied mathematics community in the context of regularization of other inverse problems, and various methods were compared systematically by Hanke & Hansen [6]. One of the most satisfying schemes involves the L-curve, suggested two decades ago by Lawson & Hanson [16], and popularized recently by Hansen [7]. In the L-curve approach one optimizes (2) for a number of values of λ and plots the resulting $r(x)$ versus $q(x)$. The result should look like an L-shaped curve. A good value of λ would be one at which the curvature of the L-curve is greatest, and one might apply a one dimensional optimization technique to determine a good λ based on this criterion.

In Section 3 we shall look more closely at the assumptions underlying the regularization techniques, and we consider the justification for (2) and the properties of L-curves in some more detail. In contrast to previous treatments we consider the case when the parameters are constrained to lie in a convex set of admissible parameters. This permits, for example, the accommodation of the nonnegativity constraints in the PET problem. (The traditional, but slow EM-algorithm handles these automatically; another alternative, used e.g. in Zhou et al. [28], is the use of additional penalty terms for negative parameters.) We also discuss approximations to the optimal (Tikhonov) L-curve that are more relevant in the context of iterative methods.

In Section 4 we present an algorithm for determining a good regularized solution based on preconditioned conjugate gradients (CG). Instead of optimizing (2) for many values of λ, we adjust the value of λ adaptively during the CG-iteration, so that the desired image can be reconstructed with a small number of iterations.

Section 5 discusses numerical results for the PET reconstruction problem. However, although the application discussed in this paper is the image reconstruction problem in positron emission tomography, the technique can be applied as well to other ill-posed or ill-conditioned problems where regularization is needed to reduce the sensitivity of the solution to the noise in the data. Finally, Section 6 summarizes our findings.

3 Regularization as a multiobjective optimization problem

In typical 'inverse' problems arising in applications, such as the PET problem, the coefficient matrix is very ill-conditioned, reflecting that the problem arises from the discretization of an ill-posed problem. Since the right-hand side is contaminated with noise from modeling errors and measurement errors, the solution is only determined up to additive corrections by near-null vectors of the matrix. Near-null vectors may have quite large components, generally of an oscillating nature, and the oscillations picked up in this way are the cause of the snowy contribution to the reconstructed picture.

To suppress these terms – in standard terminology, to *regularize* the solution, one must exploit qualitative features of the expected solution, in particular the (piecewise) continuity of the solution, and a typical approach is to demand that some linear transform of the solution (suitable differences measuring 'continuity' or 'smoothness') must kept reasonably small.

In this section, we discuss the general linear regularization problems, where, with suitable matrices A and J, the informal problem

$$Ax \approx b, \tag{11}$$

$$Jx \quad \text{well-scaled and not too large} \tag{12}$$

is formalized by requiring that both

$$r(x) := \|Ax - b\|_2^2 \tag{13}$$

and

$$q(x) := \|Jx\|_2^2 \tag{14}$$

be minimized over a convex set of admissible vectors x. However, proofs are written in such a way that they also hold for convex nonquadratic q such as (5), and qualitatively, the results tend to be true also for more general nonlinear q and r.

In general, (13) and (14) are conflicting requirements, and one can minimize one only at the expense of the other. Therefore one settles for unimprovable vectors x (also called *efficient* or *Pareto optimal*, see, e.g., SAWARAGI et al. [19]), which have the property that for all admissible vectors y, either $r(y) \geq r(x)$ or $q(y) \geq q(x)$. If we write $q_0 = q(x)$, this is equivalent to the requirement that x minimizes $r(x)$ under the constraint $q(x) \leq q_0$, and by introducing a Lagrange multiplier λ, one sees that x must be a stationary point of

$$f_\lambda(x) := r(x) + \lambda q(x) \tag{15}$$

for some $\lambda \geq 0$. Since r and q are convex quadratics, x is in fact a minimizer of $f_\lambda(x)$. Conversely, it is easy to see that all minimizers of (15) with $\lambda \geq 0$ are Pareto optimal.

Alternatively, if one minimizes a differentiable compromise function of the form $\omega(q(x), r(x))$, where ω is strictly monotone in both arguments, the gradient

$$\nabla_q \omega(q(x), r(x)) q'(x) + \nabla_r \omega(q(x), r(x)) r'(x)$$

must vanish, whence, again, x is a stationary point of (15) with

$$\lambda = \nabla_q \omega(q(x), r(x)) / \nabla_r \omega(q(x), r(x)) \geq 0.$$

Similarly, the optimization of $r(x)$ under the constraint $q(x) \leq \gamma$, or of $q(x)$ under the constraint $r(x) \leq \epsilon$, or more generally, of a compromise function $\omega(q(x), r(x))$ under a constraint $c(q(x), r(x)) \leq \gamma$ can again be shown to lead to a stationary point of (15). Thus (15) is the most general compromise function that needs to be considered, and it is the simplest since it is a convex quadratic function of x. The use of (15) to get an acceptable compromise goes back to Tikhonov [24, 25] for the case of solving ill-posed integral equations, and is generally referred to as *Tikhonov regularization*. See, e.g., Morozov [17] and Engl [2]; the latter also discusses other regularization techniques.

The choice of λ is more an art than a science; usually, solutions are computed for a large number of λ's, and one of them is selected by suitable heuristics; see e.g. Hansen [7], Hanke & Hansen [6] and Honerkamp & Weese [11]. The computational work involved is acceptable when the system is so small that a generalized singular value decomposition can be computed which reduces the linear problem to diagonal form, or when the matrices A and J have a narrow band structure so that function and gradient evaluations are cheap. However, for many problems, such as PET, the matrices are large and their sparsity structure is not sufficiently regular, which makes these approaches very expensive.

In the following, we take a closer look at the distribution of the set of all pairs $(q(x), r(x))$ and prove some general results that allow a flexible iterative approach to the regularization problem.

Definition. $Q = \{(q,r) | q, r \geq 0\}$ *denotes the positive quadrant in* $\mathbb{R}^2$. *We denote by* $\mathrm{conv}(S)$ *the* **convex hull** *of a subset* S *of* Q, *i.e., the set of all convex linear combinations* $(\sum \alpha_k q_k, \sum \alpha_k r_k), \alpha_k \geq 0, \sum \alpha_k = 1$ *of points* $(q_k, r_k) \in S$. *We also define the* **envelope** *of* S *to be the set* $\mathrm{env}(S)$ *consisting of all pairs* $(q', r') \in Q$ *such that* $q' \geq q, r' \geq r$ *for some* $(q, r) \in \mathrm{conv}(S)$.

It is easy to see that the envelope of a nonempty set $S \in Q$ is unbounded, and its boundary is given by some L-shaped curve, which we call the *L-curve* of S. The definition of an envelope immediately implies that the full envelope consists of the pairs (q, r) with $r \geq \psi(q)$, where $\psi : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$ is a convex and monotone decreasing function whose graph is the L-curve. When S is finite, the L-curve is piecewise linear and bends at vertices that are essentially the points of S lying on the L-curve. (A few may be missing if there are collinear triples of points.)

Theorem. *Suppose the rows of the matrix* (A^T, J^T) *are linearly independent. Then:*

(i) For each $\lambda > 0$, *the optimization problem* (15) *has a unique solution* x_λ.

(ii) Let $q_\lambda := q(x_\lambda)$ and $r_\lambda := r(x_\lambda)$. The curve Γ consisting of all pairs (q_λ, r_λ) is strictly monotone decreasing and convex, and its envelope contains the envelope of any set of pairs $(q(x), r(x))$. In particular, Γ is the L-curve of the envelope of the set of all pairs $(q(x), r(x))$.

Proof. We give the proof in such a way that one can see that the theorem is also valid in the presence of convex constraints, and for more general convex functionals in place of q and r such that (15) is convex and has a unique finite minimizer.

(i) Since the rows of the matrix (A^T, J^T) are linearly independent, the Hessian $2(A^TA+\lambda J^TJ)$ of (15) is positive definite when $\lambda > 0$. Hence (15) is uniformly convex and has a unique minimizer x_λ. Since the Hessian is nonsingular, the implicit function theorem implies that x_λ depends continuously differentiable on λ.

(ii) For any $x^0 \in \mathbb{R}^n$, the vector $\hat{x}$ minimizing $r(x)$ under the constraint $q(x) \le q(x^0)$ has the form $\hat{x} = x_\lambda$ for some Lagrange multiplier λ and satisfies $r(\hat{x}) \le r(x^0)$. Hence any $(q(x^0), r(x^0))$ is in the envelope of Γ.

If we restrict the optimization to the curve $\{x_\mu | \mu \ge 0\}$ we see that $r_\mu + \lambda q_\mu$ takes its minimum at $\mu = \lambda$. Therefore the derivative with respect to μ vanishes at $\mu = \lambda$, giving

$$r'_\lambda + \lambda q'_\lambda = 0. \tag{16}$$

Since $\lambda \ge 0$, this shows that r_λ is increasing when q_λ is decreasing, and conversely. Since r_λ takes its global minimum at $\lambda = 0$, it must be increasing for small λ. To show a global strict monotone decrease of the curve Γ, it therefore suffices to show that Γ contains only a single point (q, r) with fixed value of either q or r.

To see this, note that minimality gives

$$r_\lambda + \lambda q_\lambda \le r_\mu + \lambda q_\mu, \quad r_\mu + \mu q_\mu \le r_\lambda + \mu q_\lambda.$$

Now if $q_\lambda = q_\mu$ for $\lambda, \mu > 0$, this implies that $r_\lambda = r_\mu$, and conversely. Thus, the point (q_λ, r_λ) is uniquely determined by each of its components, though λ itself need not be unique. This latter situation seems very pathological, and it is likely that it never occurs. For the linear case without constraints, this can indeed be proved using a generalized singular value decomposition; see HANSEN [7].

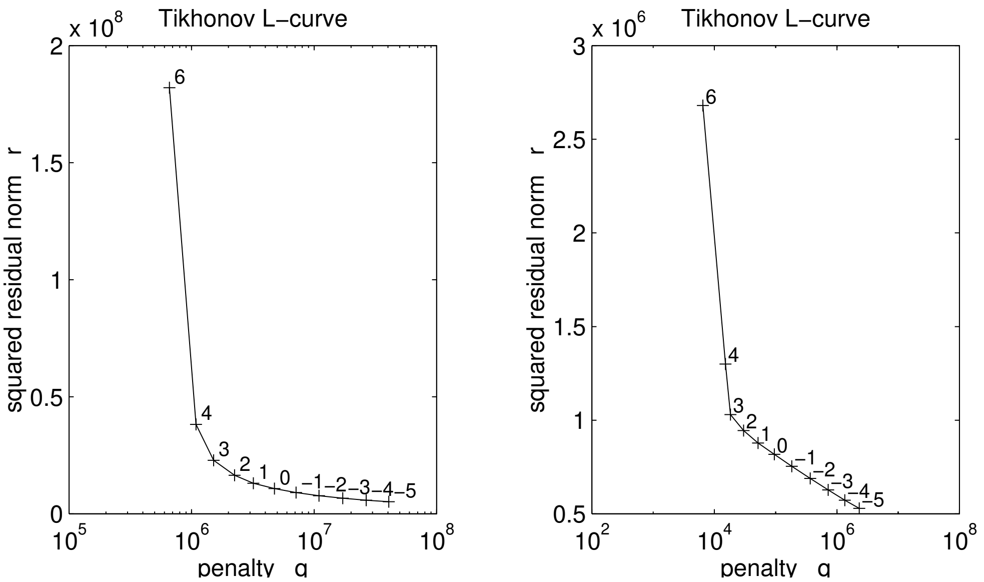

Figure 1: Approximate Tikhonov L-curves for a PET problem with 10 million (left) and 1 million (right) recorded photon pairs and quadratic penalty (4); the numbers in the figure refer to $\log_2 \lambda$.

To show that Γ is convex, let $0 \leq \mu' \leq \mu \leq \mu''$ and $\mu \leq \lambda_0$. By monotonicity, the number $\alpha := (q_\mu - q_{\mu'})/(q_{\mu''} - q_{\mu'})$ lies between 0 and 1. The convexity of q and r implies that the vector $x^0 := x_{\mu'} + \alpha(x_{\mu''} - x_{\mu'})$ satisfies $q(x^0) \leq q_{\mu'} + \alpha(q_{\mu''} - q_{\mu'}) = q_\mu$ and $r(x^0) \leq r_{\mu'} + \alpha(r_{\mu''} - r_{\mu'})$. Constructing $\hat{x}$ from x^0 as before, we see that $q_\lambda \leq q_{\mu'} + \alpha(q_{\mu''} - q_{\mu'}) = q_\mu$ and $r_\lambda \leq r_{\mu'} + \alpha(r_{\mu''} - r_{\mu'})$ for some $\lambda > 0$. The first inequality gives us $\lambda \geq \mu$, hence $r_\mu \leq r_\lambda \leq r_{\mu'} + \alpha(r_{\mu''} - r_{\mu'})$. Thus (q_μ, r_μ) lies on or below the line segment joining $(q_{\mu'}, r_{\mu'})$ and $(q_{\mu''}, r_{\mu''})$. $\square$

We call the curve Γ in (iii) the *Tikhonov L-curve*. The convention to draw the curves with the residual size on the vertical axis agrees with that used in Chapter 26 of Lawson & Hanson [16]. For points generated by an iterative method, this convention produces points 'going down', which looks natural. Note, however, that the L-curves considered by Hansen [7] are the transposes of ours, since he displays the residuals on the horizontal axis. Moreover, both [16] and [7] draw L-curves using log-log scales, and in such scales the curves no longer look convex over the full range of λ.

Example. We generated the matrix for the PET problem defined by Shepp & Vardi [20] using a 128×128 grid and 128 detectors, resulting in $\binom{128}{2}$ approximate equations for 128^2 variables; but roughly a quarter of the latter (the values outside a circle inscribed to the box defining the grid) were fixed at zero. Two right-hand sides were constructed from a phantom distribution consisting of simple geometric shapes to facilitate the interpretation of the results, by simulating 1 and 10 million random annihilations, respectively, from this distribution. Figure 2 displays the resulting densities for the annihilation points. For $q(x)$ given by the convex quadratic (4), a few points of the corresponding Tikhonov L-curve — approximated by the result of 99 conjugate gradient iterations — are drawn in Figure 1. As it should be, the curve is indeed L-shaped.

Proposition. *Let $(q_0, r_0), (q_1, r_1), \ldots, (q_N, r_N)$ be the vertices of an L-curve of a finite set S, ordered such that*

$$r_0 > r_1 > \ldots > r_N. \tag{17}$$

Then

$$q_0 < q_1 < \ldots < q_N, \tag{18}$$

and the slopes

$$s_k := (r_{k-1} - r_k)/(q_k - q_{k-1}) \quad (k = 1, \ldots, N) \tag{19}$$

satisfy the relations

$$s_1 > s_2 > \ldots > s_N > 0. \tag{20}$$

Proof. Since the L-curve is piecewise linear, monotonicity gives (18) and convexity gives (20); all inequalities are strict since, by definition of a vertex, the derivative is discontinuous at each vertex. □

When solving (15) for a particular value of λ with an iterative method, we generate a sequence of points converging to a point on the Tikhonov L-curve. If we repeat this for many values of λ, we generate a set of points whose L-curve closely approximates that of the Tikhonov L-curve. In particular, the

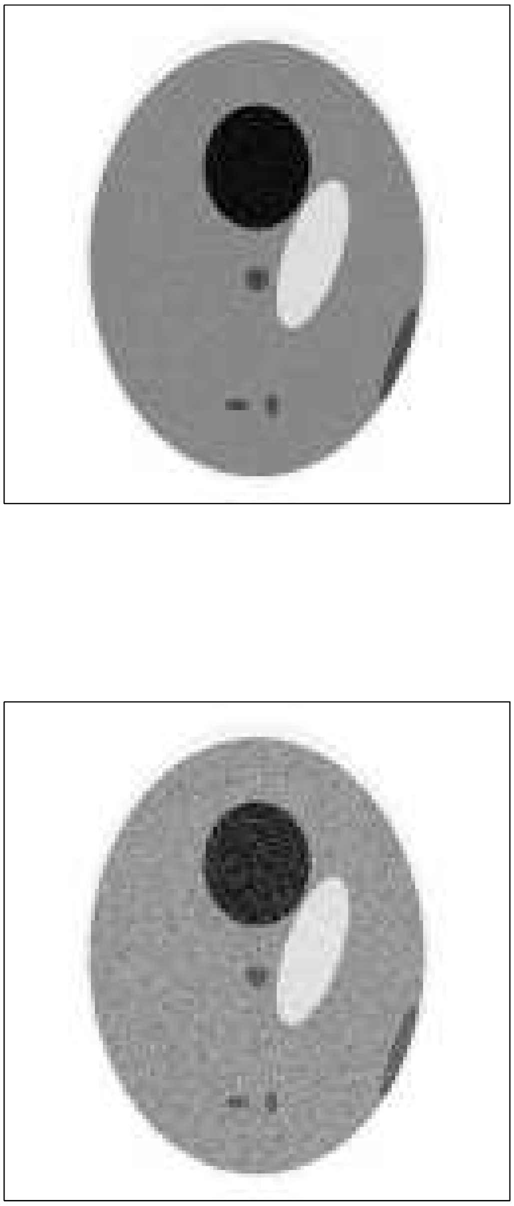

Figure 2: Phantom images from 10 million (top) and 1 million (bottom) simulated recorded photon pairs

vertices of the approximating L-curves determine at any stage of the calculation a set of currently best compromises for the regularization problem. As in the case of Tikhonov regularization, the selection of a particular one of these compromises must rely on heuristics.

Intuitively, the most appropriate compromise seems to be the vertex where the L-curve is 'most bent', cf. HANSEN [7]. However, the measure for assessing how much the curve bends at a vertex must be chosen with care. Indeed, if we simply rescale (11) and/or (12), we can change the shape of the L-curve drastically; clearly we must find a measure that is invariant under such changes. A natural further requirement is that this measure only depends on the vertex and its two neighboring vertices.

Scaling q and r simultaneously by the same amount leaves the slopes s_k defined in the above proposition unchanged, and scaling them by different amounts multiplies all s_k by the same factor. Hence the quotient of two consecutive slopes,

$$c_k := s_k / s_{k-1}, \tag{21}$$

is scale invariant, and approaches the minimal value 1 when the slopes are nearly the same, i.e., when the L-curve is hardly bent. Thus a large quotient (21) indicates a strong bend.

Another indicator results from looking at curvature terms measuring bentness based on a difference quotient of slopes, with correction factors designed to make this difference quotient scaling invariant. The simplest choice is

$$c_k := q_k(s_k - s_{k+1})/(r_{k-1} - r_{k+1}). \tag{22}$$

This expression is always positive, and approaches zero as the bend at the vertex becomes weaker and weaker. Again, a large value of (22) indicates a strong bend.

Thus we may consider the following criterion for deciding on a corner:

Definition. *The L-curve is said to be* **most bent** *at the vertex* (q_k, r_k) *(*$1 \le k \le N-1$*) where* c_k*, given by* (21) *(***version 1***) or* (22) *(***version 2***), is largest. This point is called the* **corner** *of the L-curve; it is called* proper *when* $k \ne 1, N-1$.

Remarks. 1. In many cases, the two versions define the same corner; but sometimes the decision depends on the version chosen.

2. Due to the nature of this test it is impossible to check for a bend at the leftmost and rightmost vertex ($k = 0, N$). Hence we accept a corner as proper only if it is interior enough, i.e., if its index lies between 2 and $N-2$. A corner at $k = 1, N-1$ is a clear sign that only one branch of the L-curve has been found, and further points should be computed to see whether the corner moves or becomes proper.

4 An envelope guided conjugate gradient algorithm

Hansen & O'Leary [8] suggest a scheme for determining the point of maximum curvature of the Tikhonov L-curve, by calculating suitable points along the L-curve, using a bracketing scheme to find the point of maximum curvature. Clearly, their algorithm assumes that the calculation of the optimum x of the penalty function (2) for a new value of λ is not time consuming. This assumption, unfortunately, is not reasonable for the PET problem. This does not mean that the L-curve approach must be abandoned, but it must be suitably modified.

In our iterative approach we move in the (q, r) plane trying to approximate a corner of the Tikhonov L-curve, but instead of finding points on the curve itself, we use the L-curves determined by the envelopes of the set of points generated so far to guide us to the desired destination. We try to get better and better approximations to a point of large curvature of the Tikhonov L-curve by monitoring the corners of the current L-curves.

The sets of points whose envelope defines the current L-curves are generated by the preconditioned conjugate gradient method described in Kaufman [14], applied to the objective function (2) for suitable values of λ. For each particular value of λ, our aim is not so much to optimize (2) as to try to find points outside the current envelope that will bring the L-curve closer to the optimal Tikhonov L-curve, especially near the desired corner.

Assuming one has computed points x_k with which one has determined values (q_k, r_k), one can determine the vertices of the envelope as follows:

Envelope construction algorithm.

1. Sort the points so that $r_k \le r_{k+1}$*, i.e.,* r *decreases.*

2. Discard any point k *for which* $q_k \ge q_{k+j}$ *for all* j*. Thus all points left in the set have been ordered so that* r *decreases and* q *increases. Renumber the ordered points with consecutive numbers.*

3. Insist on convexity by computing for each point k*,*

$$s_k = (r_{k-1} - r_k)/(q_k - q_{k-1});$$

discard points k *such that* $s_k \le s_{k+1}$*. If any points have been discarded, renumber the points and repeat Step 3.*

4. The points left in the set are the vertices of the envelope. The corner of the envelope is the vertex with the largest value of (21) *(version 1) or* (22) *(version 2).*

When updating the set of points by adding new points, points already discarded in a previous step cannot become new vertices; hence it is sufficient to store the vectors x_k (and the associated (q_k, r_k)) that belong to current vertices. This is an important consideration, since the set of vertices is typically rather small, often containing only 1 or 2 elements. To keep storage requirements low, we limited the number of vertices retained to 8, overwriting — if necessary — the first or last vertex, whichever is further away from the current corner.

Using the Envelope Algorithm one can consider the following generic scheme for obtaining a regularized solution. (This generic scheme is still somewhat vague, and we discuss specific details afterwards.)

Generic envelope guided algorithm.

1. Set $\lambda = 0$*.*

2. While the corner is the highest-numbered vertex of the envelope (in particular, while the envelope has a single vertex only) proceed as follows:

(a) Take a step with your favorite optimizer for (2)*.*

(b) Determine the envelope of all the points seen thus far and find its corner.

3. Eliminate all points before the corner from consideration.

4. While the corner of the envelope is near (in the tail strategy discussed in detail below: at or next to) the left endpoint of the envelope proceed as follows:

(a) Apply a few (we used 3) steps of an optimization procedure to the new f in (2)*. For each iterate add the pair (q, r) to the list of points defining the current L-curve, and update the envelope and its corner.*

(b) Adjust λ.

The algorithm terminates when a limit on the number of iterations is reached, and returns the corner of the final L-curve as approximate solution of the regularization problem.

In Step 4 (b) one would wish to choose λ to decrease both r and q. Locally, at x_k we are assured that q would decrease if

$$\nabla q(x_k)^T(-\nabla r(x_k) - \lambda \nabla q(x_k)) < 0$$

which suggests that

$$\lambda > \lambda_{min} = \max\left(\epsilon, -\frac{\nabla q(x_k)^T \nabla r(x_k)}{\nabla q(x_k)^T \nabla q(x_k)}\right), \tag{23}$$

where ϵ is the machine precision. Note that if $\lambda_{min} > \epsilon$, then at x_k, λ_{min} also minimizes

$$g(\lambda) = \|\nabla r(x_k) + \lambda \nabla q(x_k)\|^2. \tag{24}$$

Hence, whenever we are close to the Tikhonov L-curve we can expect $(q_{\lambda_{min}}, r_{\lambda_{min}})$ to be a point on the Tikhonov L-curve close to (q_k, r_k). Locally, we are assured that r would decrease if

$$\nabla r(x_k)^T(-\nabla r(x_k) - \lambda \nabla q(x_k)) < 0$$

which suggests that

$$\lambda < \lambda_{max} = -\frac{\nabla r(x_k)^T \nabla r(x_k)}{\nabla q(x_k)^T \nabla r(x_k)}. \tag{25}$$

Since the gradients of q and r are already needed to compute the gradient of the objective function, required in most optimization procedures, the cost for computing λ_{max} and λ_{min} is negligible. Note that often λ_{min} is smaller than λ_{max} by many orders of magnitude.

The most robust variant of the general scheme, which we call the *tail strategy*, is similar to the adaptive annealing algorithm in SYMONDS, HAN, SANTAGO & SNYDER [22]. One can use the envelope as a gatekeeper and begin every major iteration in Step 4(b) with x from the previous iteration. Since, in the PET application, the major features in the image are defined by the end of Step 2 but noise still slightly obscures the signal, one does not have to be that careful with the choice of λ. At the first encounter of step 4(b) we set λ to the geometric mean of λ_{min} and λ_{max} from (23) and (25), i.e. we set λ to $\hat{\lambda} = (\lambda_{max}\lambda_{min})^{1/2}$ and take 3 steps of the PCG algorithm. At subsequent encounters of Step 4(b), the current value of λ, λ_k, is reset to λ_{k+1} by the following scheme

$$\lambda_{k+1} = \begin{cases} \min(\sigma\lambda_k, .5(\lambda_k + \lambda_{max})) & \text{if } (q_k, r_k) \text{ is below the corner;} \\ \max(\lambda_k/2, .5(\lambda_k + \lambda_{min})) & \text{if } (q_k, r_k) \text{ is above the corner;} \\ \lambda_k & \text{if } (q_k, r_k) \text{ is the corner of the envelope.} \end{cases} \tag{26}$$

For the tail strategy, σ in (26) was set to 4.

A less robust strategy and one more sensitive to the choice of λ begins every major iteration in Step 4(b) with x as the corner of the last envelope if the envelope has more than two points. We call this the *corner* strategy in our graphs. With the corner strategy, the features in the data may not be that well defined at the first iteration of Step 4 and one has to be more careful that λ does not become too large too quickly. Our computational experience suggested that the first nonzero value of λ should be λ_{min} from (23). On subsequent encounters of step 4(b), the strategy of adjusting λ given in (26) was used with σ set to 2. Note that the formula for λ_{min} is based on just a first order estimate. We found that for highly nonlinear functions like (6), that after changing λ the next point the (q, r) plot was sometimes interior to the envelope because this iterate was a simple steepest descent step rather than a conjugate gradient step and the estimate of λ_{min} was too low. Our experience suggested restarting with the x corresponding to the next point in the envelope with a lower value of r was superior to trying to guess the correct λ. Termination of the whole algorithm for the corner strategy depends on whether the corner position has changed significantly. The corner strategy is much more sensitive to wrong decisions and for this reason the authors prefer the tail strategy.

Step 3 was included in the generic algorithm above to eliminate the pos-

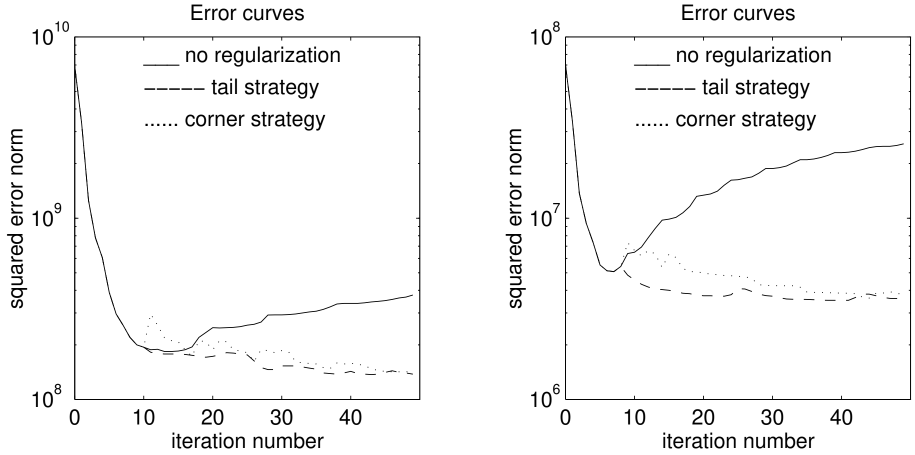

Figure 3: Sum of squares of the difference between computed images and designed image for 10 million (left) and 1 million (right) photon pairs, for the preconditioned conjugate gradients method without penalty term, and with a quadratic penalty term using the tail strategy or the corner strategy

sibility that the final corner would be at or above the first corner, a problem that had occurred in practice.

5 Numerical results

In this section we give test results for the PET problem described in the Example of Section 3, with data corresponding to Figure 2. We considered several choices of the penalty function. Unless indicated otherwise, we computed the corners using version 1, i.e., equation (21) as a measure of bentness.

For our experiments, the underlying optimization method used was the preconditioned conjugate gradient method (pcg) of KAUFMAN [14] that successfully handles nonnegativity constraints. Unless indicated otherwise, the starting vector for the first optimization was always uniform, except outside a circle inscribed to the box defining the grid, where zero entries were forced for all images; normalization was such that the total sum of entries equal to the number of photon pairs counted.

Figure 3 compares the size of the squared error norm during the iteration for both tail strategy and corner strategy with the corresponding result for the unregularized pcg method. The penalty term is the quadratic (4); but for other penalty functions, the situation is qualitatively similar. One sees that both strategies successfully avoid the growth of the error after the initial reduction that is so characteristic of the pure least squares method on ill-conditioned problems.

The occasional oscillations of the error in the tail strategy are probably due to the fact that sometimes our criterion for determining the corner is not ideal. Oscillations in the corner strategy are more pronounced, because we always restart at a corner which seems to waste some of the already available information; apparently the corner is mostly a too conservative guess for the best restart point. Both strategies reduce the error significantly below that of the best unregularized iterate, with an advantage in speed for the tail strategy.

The number of photon pairs used in the simulation determines the noise level; it decreases with an increasing number of recorded photon pairs. A number of 1 million photon pairs is probably the lower limit that still allows to recover details, and we treat this more difficult case first.

Case 1: 1 million photon pairs

Figures 4 and 5 give a more detailed view of the progress by displaying for the 1 million photon case the curves in the (q, r)-plane defined by the iterates and the (approximate) Tikhonov L-curve whose corner is the desired target. The penalty functions used were the quadratic penalty (4), the Huber penalty (9) with $\delta = 1$, the Green logcosh penalty (7) with $\delta = 2^8$, and the Hebert-Leahy log penalty (6) with $\delta = 2^3$. For the corner strategy we labeled the points before each restart at a new corner in order to facilitate the interpretation of the figures. We used a maximum of 32 pcg iterations.

The curve defined by the points of the pcg iteration is seen to be far away from the Tikhonov L-curve. Thus an optimal stop of the pcg iteration alone, as suggested in Hanke & Hansen [6], is not sufficient for an optimal reconstruction of the image. Both of our new strategies approach the Tikhonov L-curve much closer, with a visible advantage for the tail strategy.

It may be interesting to note that the values of λ used in the objective function generally were not close to the values of λ of a nearby point on the

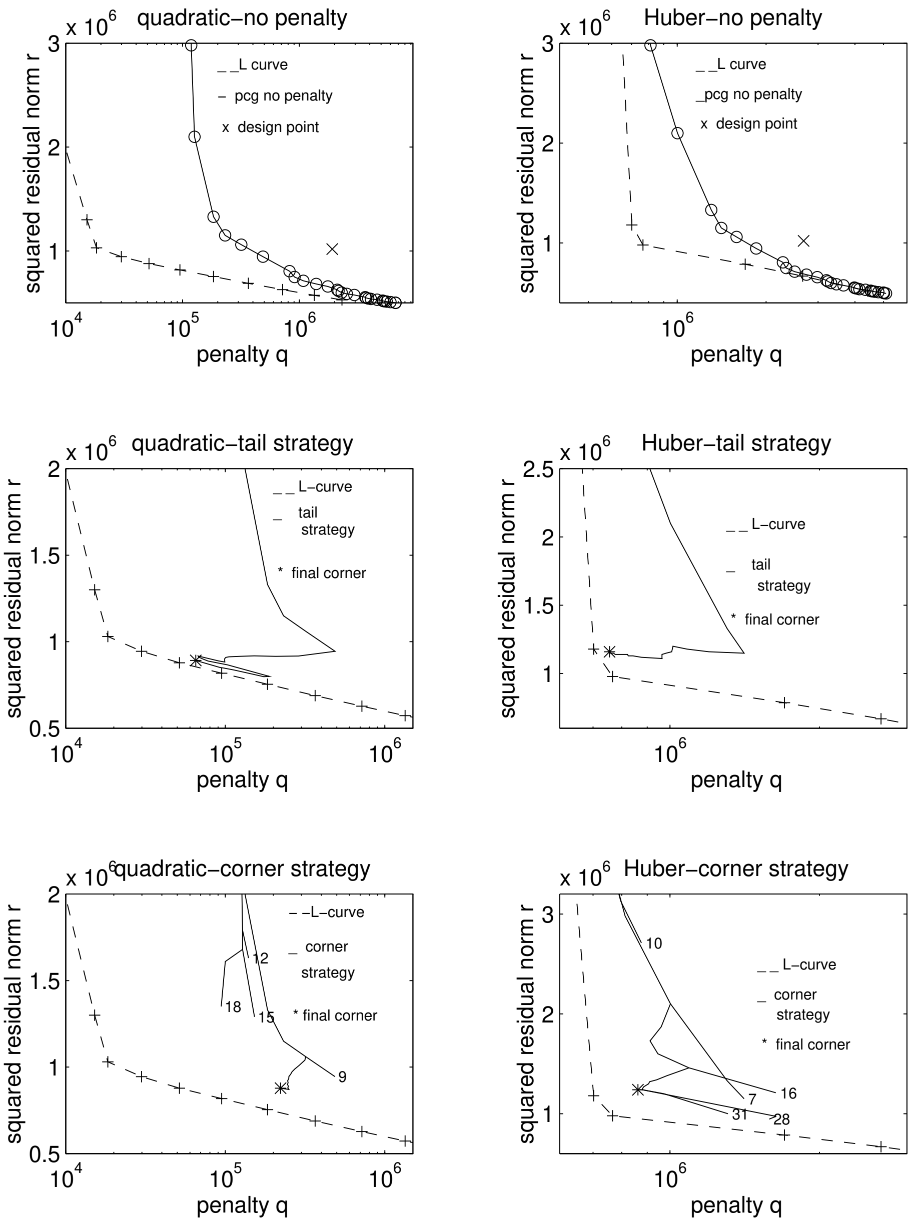

Figure 4: Tikhonov L-curve and iteration history for the quadratic (left) and Huber (right) penalty function

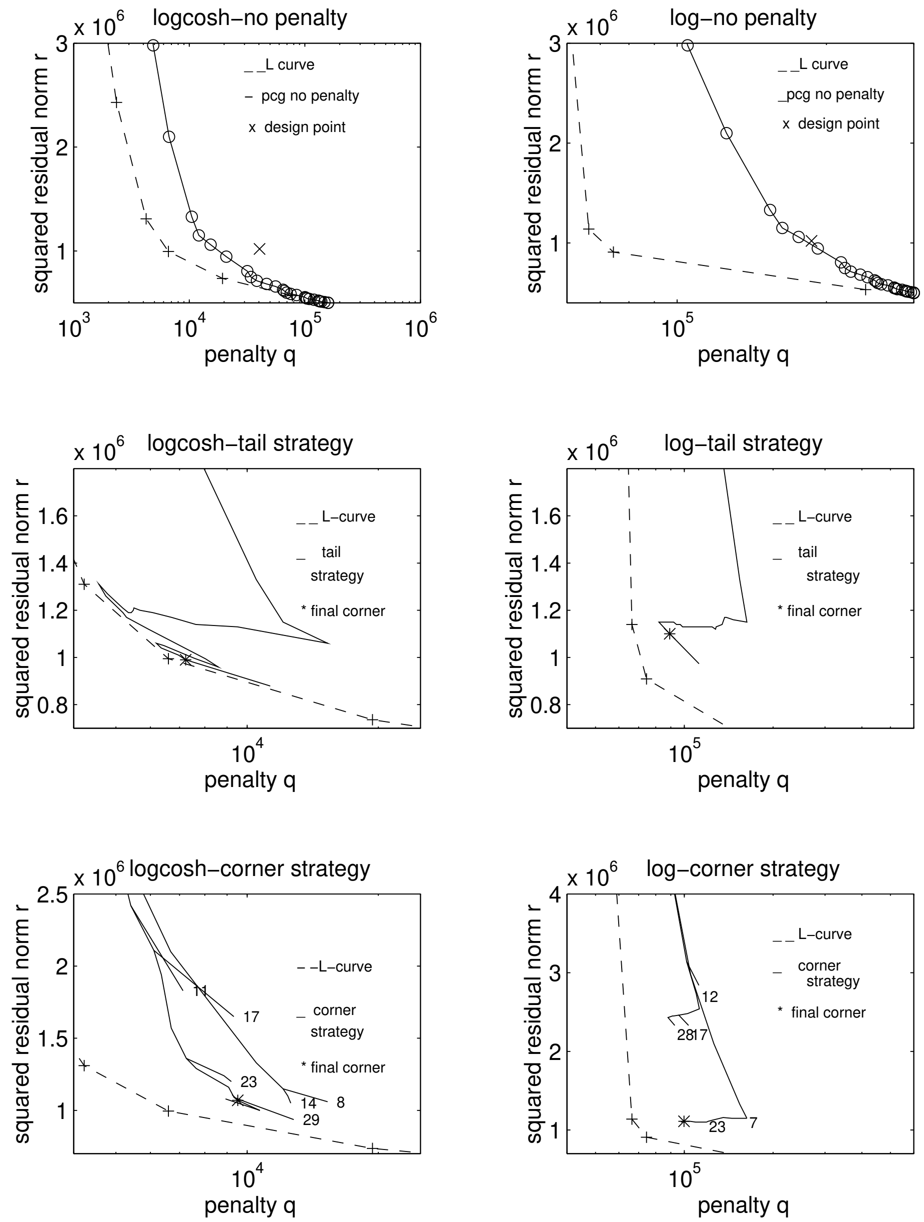

Figure 5: Tikhonov L-curve and iteration history for the logcosh (left) and log (right) penalty function

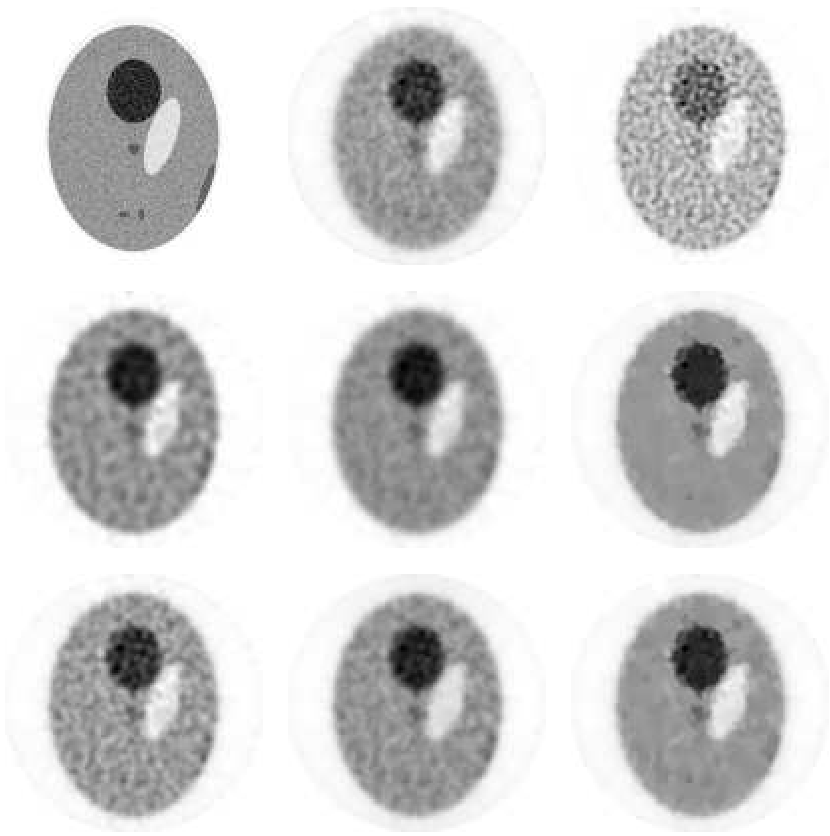

Figure 6: top row: phantom, and no penalty after 7 (best) and 15 iterations
middle row: tail strategy for quadratic, logcosh and log nonlinearity
last row: corner strategy for quadratic, logcosh and log nonlinearity

L-curve, since in our envelope strategies, the conjugate gradient method was for each λ terminated prematurely to avoid wasting iterations that would have produced points way past the desired corner.

Figure 6 shows the images computed in various ways. For the display, we transformed the components of the final vector linearly so that they span the range $[1, 256]$, and used a corresponding gray scale. The top row displays the phantom used (for comparison), the optimal approximation – with minimal Euclidean distance to the phantom – computed by the pcg method without regularization (after 7 iterations), and the deterioration resulting from continuing the unregularized iteration until iteration 15. The second row gives

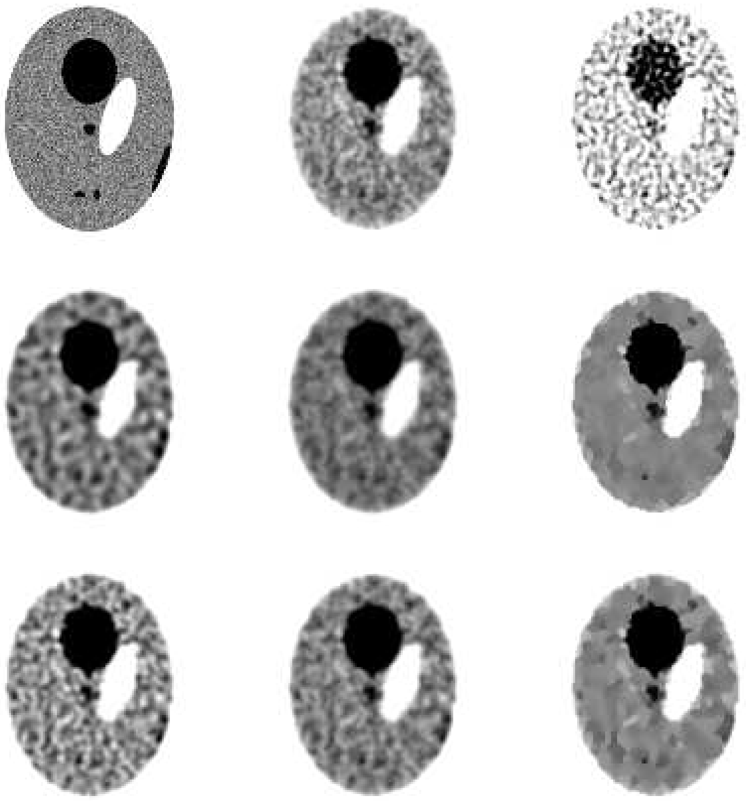

Figure 7: Same as previous figure but enhanced display

the final corner of the tail strategy (terminated after 32 iterations) for the quadratic penalty (4), the logcosh penalty (7) and the log penalty (6) with the scale parameter δ chosen as before.

Because of the noise in the phantom, the randomness in choosing the direction of the annihilating photons and the inaccuracy resulting from the discretization used to set up the matrix, the noise level only barely allows the reconstruction of the small features (representing idealized tumors).

To simulate the effect of a posteriori image processing we used the following simple enhancement technique. We replaced all pixel values < 100 by 100 and all pixel values > 200 by 200, and rescaled the new gray values to fill the interval [1,256] again. The resulting enhanced images are shown in

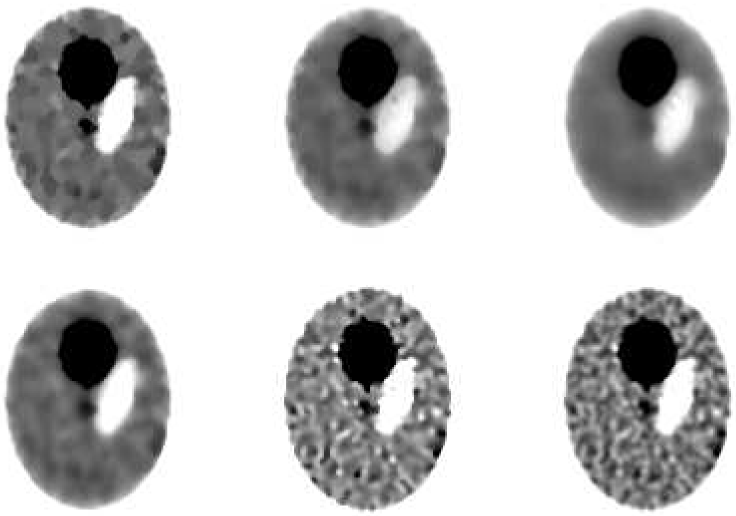

Figure 8: Sensitivity for Huber penalty (enhanced display): corner strategy for $\delta = 1, 2, 4$ (top) and $\delta = 8, 16, 32$ (bottom)

Figure 7, and display the desired details more visibly.

The results for the nonquadratic penalty functions depend on the scale parameter δ used. Unfortunately, we have not yet found a way to choose a good value for this parameter reliably in an automatic fashion. Fortunately, it turns out that the sensitivity of the results with respect to δ is not too strong when the nonlinearity is convex and asymptotically linear, i.e. for the choices (7), (8) and (9). These penalty functions, when properly scaled and unlike the redescending penalty functions (5) and (6), approach a well-defined limit penalty function as $\delta \to 0$ or $\delta \to \infty$, namely a quadratic penalty and the (nondifferentiable) absolute value penalty. However, extreme values of δ bringing the penalty close to the absolute value penalty slow down the pcg method, resulting in a degradation of the image when, as in our experiments, the number of iterations is fixed.

In Figure 8, we show, for the Huber penalty (9), the dependence of the results of the corner strategy on δ. The same is done for the tail strategy in Figure 9. Moreover, to show the influence of the formula used to compute the corner, we did these calculations also using version 2, i.e., equation (22) as a measure of bentness, giving the results displayed in Figure 10. Here

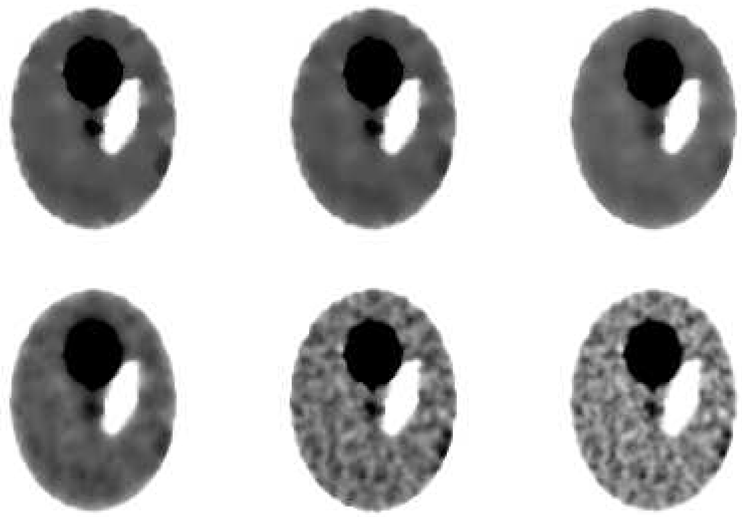

Figure 9: Sensitivity for Huber penalty (enhanced display): tail strategy for $\delta = 1, 2, 4$ (top) and $\delta = 8, 16, 32$ (bottom);

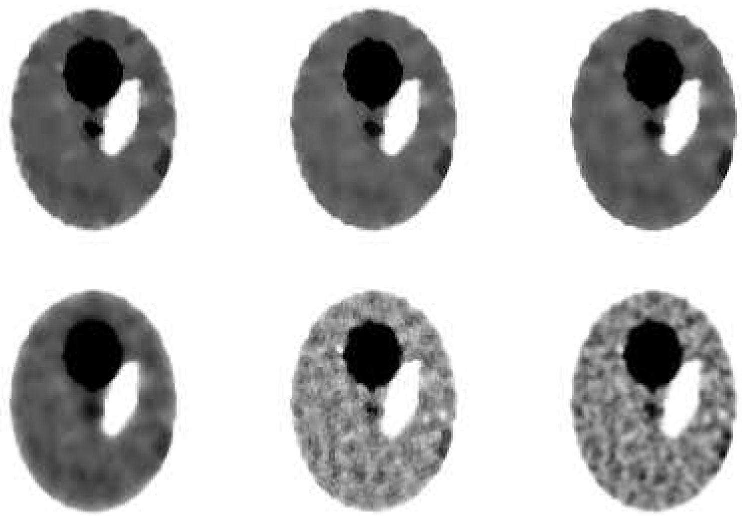

Figure 10: Same as previous figure, but with corners defined by the difference quotient

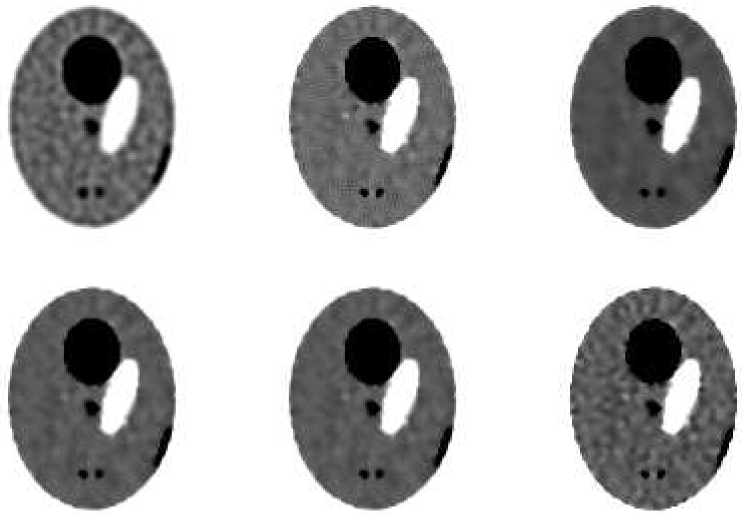

Figure 11: Tail strategy for 10 million registered photon pairs. The result of the quadratic case was used as start for the other cases; all images in enhanced display.
top row: quadratic, log for $\delta = 2^{13}$, logcosh for $\delta = 2^5$;
bottom row: Huber for, $\delta = 2^8$, multiquadric for $\delta = 2^8$, rational for $\delta = 2^{17}$.

(but not in general) version 2 produced the most pleasing results of all our methods.

Case 2: 10 million photon pairs

The high photon case is, in comparison, much easier, and produces reconstructed images of good quality. Best results were obtained using 32 iterations of the tail strategy with the quadratic penalty function (4) to compute a good starting point, used with the tail strategy for the nonlinear penalties (another 32 iterations). Figure 11 shows the results for optimal choices of the scaling parameter δ, found by trying all powers of 2 in a reasonable range.

Again the convex, asymptotically linear penalty terms removed most high frequency noise and were less sensitive to the precise choice of δ. The most insensitive formula was the multiquadric penalty (8), which produced excellent results over a wide range of δ's; see Figure 12. A hitherto unexplored

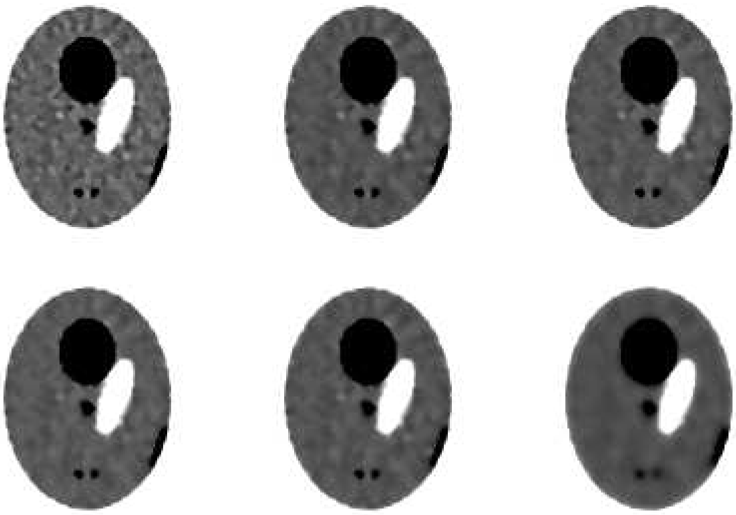

Figure 12: Tail strategy for 10 million registered photon pairs; all images in enhanced display.
multiquadric with $\delta = 2^2, 2^4, 2^6$ (top), $\delta = 2^8, 2^{10}, 2^{12}$ (bottom)

semirational formula,

$$q(x) = \sum_j \sum_{i \sim j} \frac{(x_i - x_j)^2}{|x_i - x_j| + \delta}, \tag{27}$$

behaved very similar, but avoids the calculation of square roots.

6 Conclusion

Using ideas from multiobjective optimization, we generalized the concept of an L-curve. Based on the properties of L-curves, we described several strategies for automatically adjusting the regularization parameter in iterative techniques based on Tikhonov regularization. Compared with the traditional approach of just truncating a local optimization method at a suitable iterate, the new methods try to approach the corner of the optimal (Tikhonov) L-curve.

The new techniques were successfully tested for the PET image reconstruction problem with noisy data. At low noise level (high photon count) the resulting images reproduced the test image without the introduction of artifacts, such as spurious oscillations. At high noise level (low photon count) where a truncation method only shows the gross features, the new methods were able to reproduce all features of the test image, but spurious features could not be fully suppressed.

We also made systematic tests of a number of choices for the nonlinearity in the penalty term, and found the multiquadric penalty to be the most robust one among the nonlinearities discussed in the literature.

Our techniques are also applicable to other ill-posed problems from applied mathematics.

References

[1] M. Blauer and M.D. Levine, Image Smoothing with Shape Invariance and L_1 Curvature Constraints, in: Curves and Surfaces in Computer Vision and Graphics II (1991) (M.J. Silberman and H.D. Tagare, eds.), Proc. SPIE, Vol. 1610 (1992), 299-310.

[2] H.W. Engl, Regularization Methods for the Stable Solution of Inverse Problems, Surv. Math. Ind. 3 (1993), 71-143.

[3] S. Geman and D. Geman, Stochastic relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Trans. Pattern Anal. Mach. Intelligence 6 (1984), 721-741.

[4] S. Geman and D. McClure, Bayesian Image Analysis: An Application to Single Photon Emission Tomography, Proc. Statist. Comput. Sect., Amer. Statist. Assoc, Washington, D.C., (1985), 12-18.

[5] P. Green, Bayesian Reconstructions from Emission Tomography Using a Modified EM Algorithm, IEEE Trans. Med. Imaging 9 (1990), 84-93.

[6] M. Hanke and P.C. Hansen, Regularization Methods for Large-Scale Problems, Surv. Math. Ind. 3 (1993), 253-315.

[7] P.C. Hansen, Analysis of Discrete Ill-Posed Problems by Means of the L-Curve, SIAM Review 34 (1992), 561-580.

[8] P.C. Hansen and D.P. O'Leary, The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems, SIAM J. Sci. Computing 14 (1993), 1487-1506.

[9] T. Hebert and R. Leahy, A Generalized EM Algorithm for the 3-D Bayesian Reconstruction from Poisson Data Using Gibbs Priors, IEEE Trans. Med. Imaging 8 (1990), 194-202.

[10] T. Hebert and R. Leahy, Statistic-Based MAP Image Reconstruction from Poisson Data Using Gibbs Priors, IEEE Trans. Sign. Proc. 40 (1992), 2290-2303.

[11] J. Honerkamp and J. Weese, Tikhonov Regularization Methods for Ill-Posed Problems, Continuum Mech. Thermodyn. 2 (1990), 17-30.

[12] P.J. Huber, Robust Statistics, Wiley, New York 1981.

[13] L. Kaufman, Implementing and Accelerating the EM Algorithm for Positron Emission Tomography, IEEE Trans. Med. Imaging 6, (1987), 37-51.

[14] L. Kaufman, Maximum Likelihood, Least Squares, and Penalized Least Squares for PET, IEEE Trans. Med. Imaging 12 (1993), 200-214.

[15] D.S. Lalush and B.M.W. Tsui, Simulation Evaluation of Gibbs Prior Distributions for Use in Maximum A Posteriori SPECT Reconstructions, IEEE Trans. Med. Imaging 11 (1992), 267-275.

[16] C.L. Lawson and R.J. Hanson, Solving Least Squares Problems. Prentice-Hall, Englewood Cliffs, N.J., 1974.

[17] V.A. Morozov, Methods for Solving Incorrectly Posed Problems. Springer, Berlin 1984.

[18] S.J. Reeves and R.M. Mersereau, Optimal Regularization Parameter Estimation for Image Restauration, In: Image Processing algorithms and Techniques II, Proc. SPIE, Vol. 1452 (1991), 127-138.

[19] Y. Sawaragi, H. Nakayama and T. Tanino, Theory of Multiobjective Optimization. Acad. Press, Orlando 1985.

[20] L. Shepp and Y. Vardi, Maximum Likelihood Reconstruction in Positron Emission Tomography, IEEE Trans. Med. Imaging 1 (1982), 113-122.

[21] R.L. Stevenson, B.E. Schmitz and E.J. Delp, Discontinuity Preserving Regularization of Inverse Visual Problems, IEEE Trans. Systems, Man, Cybernetics 24 (1994), 455-469.

[22] M. Symonds, Y.-S. Han, P. Santago and W. Snyder, Reconstruction of Positron Emission Tomography Images by Using Maximum a Posteriori and Mean Field Annealing Techniques, In: Physics of Medical Imaging (1994), Proc. SPIE, Vol. 2613 (1994), 212-222.

[23] D. Terzopoulos, Regularization of Inverse Visual Problems Involving Discontinuities, IEEE Trans. Pattern Anal. Mach. Intelligence 8 (1986), 413-424.

[24] A.N. Tikhonov, Solution of Incorrectly Formulated Problems and the Regularization Method. Soviet Math. Dokl. 4 (1963), 1035-1038.

[25] A.N. Tikhonov, Regularization of Incorrectly Posed Problems. Soviet Math. Dokl. 4 (1963), 1624-1627.

[26] Y. Vardi, L. Shepp and L. Kaufman, A Statistical Method for Positron Emission Tomography (with comments by various others), J. Amer. Statist. Assoc. 80 (1985), 8-37.

[27] C.R. Vogel and M.E. Oman, Iterative Methods for Total Variation Denoising Manuscript, 1994. (umsfcvog@euler.oscs.montana.edu)

[28] Z. Zhou, R.M. Leahy and E.U. Mumcuoglu, A Comparative Study of the Effects of Using Anatomical Priors in PET Reconstruction, IEEE Trans. Med. Imaging, to appear.