The L-curve and its use in the numerical treatment of inverse problems

P. C. Hansen

Department of Mathematical Modelling,
Technical University of Denmark,
DK-2800 Lyngby, Denmark

Abstract

The L-curve is a log-log plot of the norm of a regularized solution versus the norm of the corresponding residual norm. It is a convenient graphical tool for displaying the trade-off between the size of a regularized solution and its fit to the given data, as the regularization parameter varies. The L-curve thus gives insight into the regularizing properties of the underlying regularization method, and it is an aid in choosing an appropriate regularization parameter for the given data. In this chapter we summarize the main properties of the L-curve, and demonstrate by examples its usefulness and its limitations both as an analysis tool and as a method for choosing the regularization parameter.

1 Introduction

Practically all regularization methods for computing stable solutions to inverse problems involve a trade-off between the "size" of the regularized solution and the quality of the fit that it provides to the given data. What distinguishes the various regularization methods is how they measure these quantities, and how they decide on the optimal trade-off between the two quantities. For example, given the discrete linear least-squares problem $\min \|A\,x - b\|_2$ (which specializes to $A\,x = b$ if A is square), the classical regularization method developed independently by Phillips [31] and Tikhonov [35] (but usually referred to as Tikhonov regularization) amounts — in its most general form — to solving the minimization problem

$$x_\lambda = \arg\min \left\{ \|A\,x - b\|_2^2 + \lambda^2 \|L\,(x - x_0)\|_2^2 \right\}, \tag{1}$$

where λ is a real regularization parameter that must be chosen by the user. Here, the "size" of the regularized solution is measured by the norm $\|L\,(x - x_0)\|_2$, while the fit is measured by the 2-norm $\|A\,x - b\|_2$ of the residual vector. The vector x_0 is an *a priori* estimate of x which is set to zero when no *a priori* information is available. The problem is in standard form if $L = I$, the identity matrix.

The Tikhonov solution x_λ is formally given as the solution to the "regularized normal equations"

$$\left(A^T A + \lambda^2 L^T L\right) x_\lambda = A^T b + \lambda^2 L^T L\, x_0. \tag{2}$$

However, the best way to solve (1) numerically is to treat it as a least squares problem

$$x_\lambda = \arg\min \left\| \begin{pmatrix} A \\ \lambda\, L \end{pmatrix} x - \begin{pmatrix} b \\ \lambda\, L\, x_0 \end{pmatrix} \right\|_2 . \tag{3}$$

Regularization is necessary when solving inverse problems because the "naive" least squares solution, formally given by $x_{\mathrm{LS}} = A^\dagger b$, is completely dominated by contributions from data errors and rounding errors. By adding regularization we are able to damp these contributions and keep the norm $\|L\,(x - x_0)\|_2$ of reasonable size. This philosophy underlies Tikhonov regularization and most other regularization methods. Various issues in choosing the matrix L are discussed in [4], [30], and Section 4.3 in [21].

Note that if too much regularization, or damping, is imposed on the solution, then it will not fit the given data b properly and the residual $\|A\,x_\lambda - b\|_2$ will be too large. On the other hand, if too little regularization is imposed then the fit will be good but the solution will be dominated by the contributions from the data errors, and hence $\|L\,(x_\lambda - x_0)\|_2$ will be too large. Figure 1 illustrates this point for Tikhonov regularization.

Having realized the important roles played by the norms of the solution and the residual, it is quite natural to plot these two quantities versus each other, i.e., as a curve

$$\left(\|A\,x_\lambda - b\|_2\,,\ \|L\,(x_\lambda - x_0)\|_2 \right)$$

parametrized by the regularization parameter. This is precisely the L-curve; see Fig. 2 for an example with $L = I$ and $x_0 = 0$. Hence, the

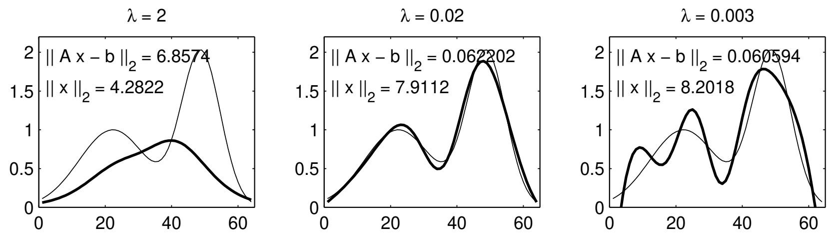

Figure 1: The exact solution (thin lines) and Tikhonov regularized solutions x_λ (thick lines) for three values of λ corresponding to over-smoothing, appropriate smoothing, and under-smoothing.

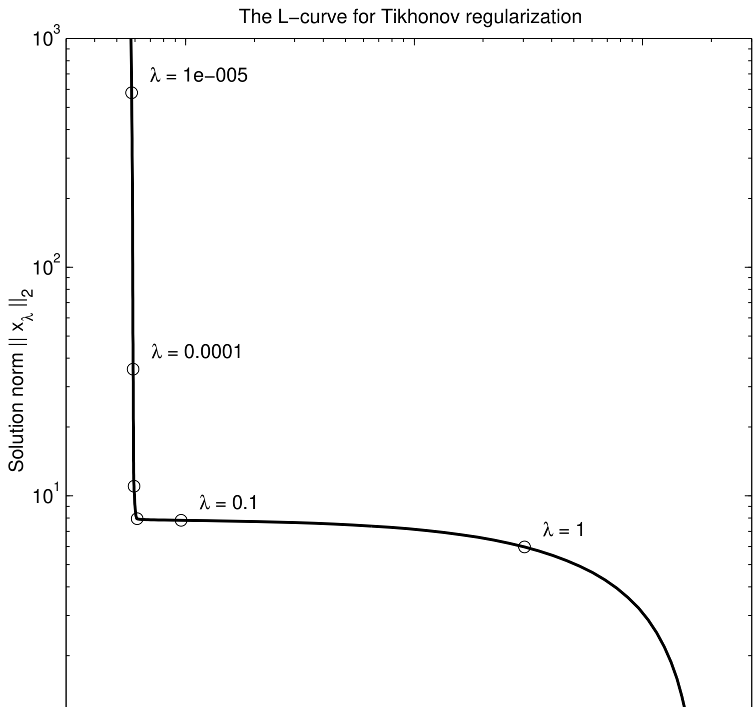

Figure 2: The generic L-curve for standard-form Tikhonov regularization with $x_0 = 0$; the points marked by the circles correspond to the regularization parameters $\lambda = 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ and 1.

L-curve is really a tradeoff-curve between two quantities that both should be controlled. Such trade-off curves are common in the applied mathematics and engineering literature, and plots similar to the L-curve have appeared over the years throughout the literature. Early references that make use of L-curve plots are Lawson and Hanson [26] and Miller [29].

Some interesting questions are related to the L-curve. What are the properties of the L-curve? What information can be extracted from the L-curve about the problem, the regularization algorithm, the regularized solutions, and the choice of the regularization parameter? In particular, we shall present a recent method for choosing the regularization parameter λ, known as the *L-curve criterion*, and discuss its practical use. This method has been used successfully in a number of applications, such as continuation problems [2], geoscience [3], and tomography [25].

The purpose of this chapter is to provide insight that helps to answer the above questions. Throughout the chapter we focus on Tikhonov regularization (although the L-curve exists for other methods as well), and we start in Section 2 with a historical perspective of Tikhonov's method. In Section 3 we introduce our main analysis tool, the singular value decomposition (SVD). In Sections 4, 5, and 6 we present various properties of the L-curve that explain its characteristic L-shape. Next, in Section 7 we describe the L-curve criterion for choosing the regularization parameter, which amounts to locating the "corner" of the L-curve. Finally, in Section 8 we describe some limitations of the L-curve criterion that must be considered when using this parameter choice method.

2 Tikhonov regularization

Perhaps the first author to describe a scheme that is equivalent to Tikhonov regularization was James Riley who, in his paper [34] from 1955, considered ill-conditioned systems $A\,x = b$ with a symmetric positive (semi)definite coefficient matrix, and proposed to solve instead the system $(A + \alpha\, I)\, x = b$, where α is a small positive constant. In the same paper, Riley also suggested an iterative scheme which is now known as iterated Tikhonov regularization, cf. §5.1.5 in [21].

The first paper devoted to more general problems was published by D. L. Phillips [31] in 1962. In this paper A is a square matrix obtained from a first-kind Fredholm integral equation by means of a quadrature rule, and L is the tridiagonal matrix $\mathrm{tridiag}(1, -2, 1)$.

Phillips arrives at the formulation in (1) but without matrix notation, and then proposes to compute the regularized solution as $x_\lambda = (A+\lambda^2 A^{-T} L^T L)^{-1} b$, using our notation. It is not clear whether Phillips computed A^{-1} explicitly, but he did not recognize (1) as a least squares problem.

In his 1963 paper [36], S. Twomey reformulated Phillips' expression for x_λ via the "regularized normal equations" (2) and obtained the well-known expression $x_\lambda = \left(A^T A + \lambda^2 L^T L\right)^{-1} A^T b$, still with $L = \mathrm{tridiag}(1, -2, 1)$. He also proposed to include the *a priori* estimate x_0, but only in connection with the choice $L = I$ (the identity matrix), leading to the formula $x_\lambda = \left(A^T A + \lambda^2 I\right)^{-1} \left(A^T b + \lambda^2 x_0\right)$.

A. N. Tikhonov's paper [35] from 1963 is formulated in a much more general setting: he considered the problem $\mathcal{K} f = g$ where f and g are functions and $\mathcal{K}$ is an integral operator. Tikhonov proposed the formulation $f_\lambda = \arg\min \left\{ \|\mathcal{K} f - g\|_2^2 + \lambda^2 \, \Omega(f) \right\}$ with the particular functional

$$\Omega(f) = \int_a^b \left(v(s) \, f(s)^2 + w(s) \, f'(s)^2 \right) ds,$$

where v and w are positive weight functions. Turning to computations, Tikhonov used the midpoint quadrature rule to arrive at the problem $\min \left\{ \|A\, x - b\|_2^2 + \lambda^2 \left(\|D_v^{1/2} \, x\|_2^2 + \|L \, D_w^{1/2} \, x\|_2^2 \right) \right\}$, in which D_v and D_w are diagonal weight matrices corresponding to v and w, and $L = \mathrm{bidiag}(-1, 1)$. Via the "regularized normal equations" he then derived the expression $x_\lambda = \left(A^T A + \lambda^2 (D_v + L^T D_w \, L)\right)^{-1} A^T b$.

In 1965 Gene H. Golub [9] was the first to propose a modern approach to solving (1) via the least squares formulation (3) and QR factorization of the associated coefficient matrix. Golub proposed this approach in connection with Riley's iterative scheme, which includes the computation of x_λ as the first step. G. Ribiere [33] also proposed the QR-based approach to computing x_λ in 1967.

In 1970, Joel Franklin [6] derived the "regularized normal equation" formulation of Tikhonov regularization in a stochastic setting. Here, the residual vector is weighted by the Cholesky factor of the covariance matrix for the perturbations in b, and the matrix $\lambda^2 L^T L$ represents the inverse of the covariance matrix for the solution, considered as a stochastic variable.

Finally, it should be mentioned that Franklin [7] in 1978, in connection with symmetric positive (semi)definite matrices A and

B, proposed the variant $x_\lambda = (A + \alpha\, B)^{-1} b$, where α is a positive scalar — which nicely connects back to Riley's 1955 paper.

In the statistical literature, Tikhonov regularization is known as ridge regression and seems to date back to the papers [23], [24] from 1970 by Hoerl and Kennard. Marquardt [27] used this setting as the basis for an analysis of his iterative algorithm from 1963 for solving nonlinear least squares problems [28], and which incorporates standard-form Tikhonov regularization in each step.

The most efficient way to compute Tikhonov solutions x_λ for a range of regularization parameters λ (which is almost always the case in practice) is by means of the bidiagonalization algorithm due to Lars Eldén [5], combined with a transformation to standard form if $L \neq I$. Several iterative algorithms have been developed recently for large-scale problems, e.g., [8], [11], [14], but we shall not go into any of the algorithmic details here.

3 The singular value decomposition

The purpose of this section is to derive and explain various expressions that lead to an understanding of the features of the L-curve for Tikhonov regularization. To simplify our analysis considerably, we assume throughout the rest of this chapter that the matrix L is the identity matrix. If this is not the case, a problem with a general $L \neq I$ can always be brought into standard form with $L = I$; see, e.g., [5] and Section 2.3 in [21] for details and algorithms. Alternatively, the analysis with a general matrix L can be cast in terms of the generalized SVD, cf. Sections 2.1.2 and 4.6 in [21]. We will also assume that the *a priori* estimate is zero, i.e., $x_0 = 0$.

Our main analysis tool throughout the chapter is the singular value decomposition (SVD) of the matrix A, which is a decomposition of a general $m \times n$ matrix A with $m \geq n$ of the form

$$A = \sum_{i=1}^{n} u_i\, \sigma_i\, v_i^T, \tag{4}$$

where the left and right singular vectors u_i and v_i are orthonormal, i.e., $u_i^T u_j = v_i^T v_j = \delta_{ij}$, and the singular values σ_i are nonnegative quantities which appear in non-decreasing order,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

For matrices A arising from the discretization of inverse problems, the singular values decay gradually to zero, and the number of sign changes in the singular vectors tends to increase as i increases. Hence, the smaller the singular value σ_i, the more oscillatory the corresponding singular vectors u_i and v_i appear, cf. Section 2.1 in [21]

If we insert the SVD into the least squares formulation (3) then it is straightforward to show that the Tikhonov solution is given by

$$x_\lambda = \sum_{i=1}^{n} f_i \, \frac{u_i^T b}{\sigma_i} \, v_i, \tag{5}$$

where $f_1, \ldots, f_n$ are the Tikhonov filter factors, which depend on σ_i and λ as

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \simeq \begin{cases} 1, & \sigma_i \gg \lambda \\ \sigma_i^2/\lambda^2, & \sigma_i \ll \lambda. \end{cases} \tag{6}$$

In particular, the "naive" least squares solution $x_{\rm LS}$ is given by (5) with $\lambda = 0$ and all filter factors equal to one. Hence, comparing x_λ with $x_{\rm LS}$ we see that the filter factors practically filter out the contributions to x_λ corresponding to the small singular values, while they leave the SVD components corresponding to large singular values almost unaffected. Moreover, the damping sets in for $\sigma_i \simeq \lambda$.

The residual vector corresponding to x_λ, which characterizes the misfit, is given in terms of the SVD by

$$b - A\,x_\lambda = \sum_{i=1}^{n} (1 - f_i)\, u_i^T b\, u_i + b_0, \tag{7}$$

in which the vector $b_0 = b - \sum_{i=1}^{n} u_i\, u_i^T b$ is the component of b that lies outside the range of A (and therefore cannot be "reached" by any linear combination of the columns of A), and $1 - f_i = \lambda^2/(\sigma_i^2 + \lambda^2)$. Note that $b_0 = 0$ when $m = n$. From (7) we see that filter factors close to one diminish the corresponding SVD components of the residual considerably, while small filter factors leave the corresponding residual components practically unaffected.

Equipped with the two expressions (5) and (7) we can now write the solution and residual norms in terms of the SVD:

$$\|x_\lambda\|_2^2 = \sum_{i=1}^{n} \left(f_i \, \frac{u_i^T b}{\sigma_i} \right)^2 \tag{8}$$

$$\|A\,x_\lambda - b\|_2^2 = \sum_{i=1}^{n} \Big((1 - f_i)\, u_i^T b \Big)^2 . \tag{9}$$

These expressions form the basis for our analysis of the L-curve.

4 SVD analysis

Throughout this chapter we shall assume that the errors in the given problem $\min \|A\,x - b\|_2$ are restricted to the right-hand side, such that the given data b can be written as

$$b = \overline{b} + e, \qquad \overline{b} = A\,\overline{x},$$

where $\overline{b}$ represents the exact unperturbed data, $\overline{x} = A^\dagger \overline{b}$ represents the exact solution, and the vector e represents the errors in the data. Then the Tikhonov solution can be written as $x_\lambda = \overline{x}_\lambda + x_\lambda^e$, where $\overline{x}_\lambda = (A^T A + \lambda^2 I)^{-1} A^T \overline{b}$ is the regularized version of the exact solution $\overline{x}$, and $x_\lambda^e = (A^T A + \lambda^2 I)^{-1} A^T e$ is the "solution vector" obtained by applying Tikhonov regularization to the pure noise component e of the right-hand side.

Consider first the L-curve corresponding to the exact data $\overline{b}$. At this stage, it is necessary to make the following assumption which is called the Discrete Picard Condition:

$$\boxed{\text{The exact SVD coefficients } |u_i^T \overline{b}| \text{ decay faster than the } \sigma_i.}$$

This condition ensures that the least squares solution $\overline{x} = A^\dagger \overline{b}$ to the unperturbed problem does not have a large norm, because the exact solution coefficients $|v_i^T \overline{x}| = |u_i^T \overline{b}/\sigma_i|$ also decay. This Discrete Picard Condition thus ensures that there exists a physically meaningful solution to the underlying inverse problem, and it also ensures that the solution can be approximated by a regularized solution (provided that a suitable regularization parameter can be found). Details about the condition can be found in [17] and Section 4.5 in [21].

Assume now that the regularization parameter λ lies somewhere between σ_1 and σ_n, such that we have both some small filter factors f_i (6) and some filter factors close to one. Moreover, let k denote the number of filter factors close to one. Then it is easy to see from (6) that k and λ are related by the expression $\lambda \simeq \sigma_k$. A more thorough analysis is given in [15]. It then follows from (8) that

$$\|\overline{x}_\lambda\|_2^2 \simeq \sum_{i=1}^{k} \left(v_i^T \overline{x}\right)^2 \simeq \sum_{i=1}^{n} \left(v_i^T \overline{x}\right)^2 = \|\overline{x}\|_2^2, \qquad (10)$$

where we have used the fact that the coefficients $|v_i^T \overline{x}|$ decay such that the last $n - k$ terms contribute very little to the sum. The expression in (10) holds as long as λ is not too large. As $\lambda \to \infty$ (and

$k \to 0$) we have $\overline{x}_\lambda \to 0$ and thus $\|\overline{x}_\lambda\|_2 \to 0$. On the other hand, as $\lambda \to 0$ we have $\overline{x}_\lambda \to \overline{x}$ and thus $\|\overline{x}_\lambda\|_2 \to \|\overline{x}\|_2$.

The residual corresponding to $\overline{x}_\lambda$ satisfies

$$\|A\,\overline{x}_\lambda - \overline{b}\|_2^2 \simeq \sum_{i=k}^{n} \left(u_i^T \overline{b}\right)^2 , \tag{11}$$

showing that this residual norm increases steadily from $\|\overline{b}_0\|_2 = 0$ to $\|\overline{b}\|_2$ as λ increases (because an increasing number of terms k is included in the sum). Hence, the L-curve for the unperturbed problem is an overall flat curve at $\|\overline{x}_\lambda\|_2 \simeq \|\overline{x}\|_2$, except for large values of the residual norm $\|A\,\overline{x}_\lambda - \overline{b}\|_2$ where the curve approaches the abscissa axis.

Next we consider an L-curve corresponding to a right-hand side consisting of pure noise e. We assume that the noise is "white", i.e., the covariance matrix for e is a scalar times the identity matrix — if this is not the case, one should preferably scale the problem such that the scaled problem satisfies this requirement. This assumption implies that the expected values of the SVD coefficients $u_i^T e$ are independent of i,

$$\mathcal{E}\left((u_i^T b)^2\right) = \epsilon^2, \qquad i = 1, \ldots, m,$$

which means that the noise component e does not satisfy the discrete Picard condition.

Consider now the vector $x_\lambda^e = (A^T A + \lambda^2 I)^{-1} A^T e$. Concerning the norm of x_λ^e we obtain

$$\begin{aligned}
\|x_\lambda^e\|_2^2 &\simeq \sum_{i=1}^{n} \left(\frac{\sigma_i\,\epsilon}{\sigma_i^2 + \lambda^2}\right)^2 \simeq \sum_{i=1}^{k} \left(\frac{\epsilon}{\sigma_i}\right)^2 + \sum_{i=k+1}^{n} \left(\frac{\sigma_i\,\epsilon}{\lambda^2}\right)^2 \\
&= \epsilon^2 \left(\sum_{i=1}^{k} \sigma_i^{-2} + \lambda^{-4} \sum_{i=k+1}^{n} \sigma_i^2\right).
\end{aligned}$$

The first sum $\sum_{i=1}^{k} \sigma_i^{-2}$ in this expression is dominated by $\sigma_k^{-2} \simeq \lambda^{-2}$ while the second sum $\sum_{i=k+1}^{n} \sigma_i^2$ is dominated by $\sigma_{k+1}^2 \simeq \lambda^2$, and thus we obtain the approximate expression

$$\|x_\lambda^e\|_2 \simeq c_\lambda\,\epsilon\,/\,\lambda,$$

where c_λ is a quantity that varies slowly with λ. Hence, we see that the norm of x_λ^e increases monotonically from 0 as λ decreases, until it reaches the value $\|A^\dagger e\|_2 \simeq \epsilon\,\|A^\dagger\|_\mathrm{F}$ for $\lambda = 0$.

The norm of the corresponding residual satisfies

$$\|A\, x_\lambda^e - b\|_2^2 \simeq \sum_{i=k}^{m} \epsilon^2 = (m-k)\,\epsilon^2 .$$

Hence, $\|A\, x_\lambda^e - e\|_2 \simeq \sqrt{m-k}\,\epsilon$ is a slowly varying function of λ lying in the range from $\sqrt{m-n}\,\epsilon$ to $\|e\|_2 \simeq \sqrt{m}\,\epsilon$. The L-curve corresponding to e is therefore an overall very steep curve located slightly left of $\|A\, x_\lambda^e - e\|_2 \simeq \|e\|_2$, except for small values of λ where it approaches the ordinate axis.

Finally we consider the L-curve corresponding to the noisy right-hand side $b = \overline{b} + e$. Depending on λ, it is either the noise-free components $u_i^T \overline{b}$ or the pure-noise components $u_i^T e$ that dominate, and the resulting L-curve therefore essentially consists of one "leg" from the unperturbed L-curve and one "leg" from the pure-noise L-curve. For small values of λ it is the pure-noise L-curve that dominates because x_λ is dominated by x_λ^e, and for large values of λ where x_λ is dominated by $\overline{x}_\lambda$ it is the unperturbed L-curve that dominates. Somewhere in between, there is a range of λ-values that correspond to a transition between the two domination L-curves.

We emphasize that the above discussion is valid only when the L-curve is plotted in log-log scale. In linear scale, the L-curve is always convex, independently of the right-hand side (see, e.g., Theorem 4.6.1 in [21]). The logarithmic scale, on the other hand, emphasizes the difference between the L-curves for an exact right-hand side $\overline{b}$ and for pure noise e, and it also emphasizes the two different parts of the L-curve for a noisy right-hand side $b = \overline{b} + e$. These issues are discussed in detail in [22].

5 The curvature of the L-curve

As we shall see in the next two sections, the curvature of the L-curve plays an important role in the understanding and use of the L-curve. In this section we shall therefore derive a convenient expression for this curvature. Let

$$\eta = \|x_\lambda\|_2^2, \qquad \rho = \|A\, x_\lambda - b\|_2^2 \tag{12}$$

and

$$\widehat{\eta} = \log \eta, \qquad \widehat{\rho} = \log \rho \tag{13}$$

such that the L-curve is a plot of $\widehat{\eta}/2$ versus $\widehat{\rho}/2$, and recall that $\widehat{\eta}$ and $\widehat{\rho}$ are functions of λ. Moreover, let $\widehat{\eta}'$, $\widehat{\rho}'$, $\widehat{\eta}''$, and $\widehat{\rho}''$ denote the first and second derivatives of $\widehat{\eta}$ and $\widehat{\rho}$ with respect to λ. Then the curvature κ of the L-curve, as a function of λ, is given by

$$\kappa = 2 \frac{\widehat{\rho}' \, \widehat{\eta}'' - \widehat{\rho}'' \, \widehat{\eta}'}{((\widehat{\rho}')^2 + (\widehat{\eta}')^2)^{3/2}}. \tag{14}$$

The goal is now to derive a more insightful expression for κ. Our analysis is similar to that of Hanke [13] and Vogel [37], but their details are omitted. Moreover, they differentiated $\widehat{\eta}$ and $\widehat{\rho}$ with respect to λ^2 instead of λ (as we do); hence their formulas are different from ours, although they lead to the same value of κ.

The first step is to derive expressions for the derivatives of $\widehat{\eta}$ and $\widehat{\rho}$ with respect to λ. These derivates are called the logarithmic derivatives of η and ρ, and they are given by

$$\widehat{\eta}' = \frac{\eta'}{\eta} \qquad \text{and} \qquad \widehat{\rho}' = \frac{\rho'}{\rho}.$$

The derivates of η and ρ, in turn, are given by

$$\eta' = -\frac{4}{\lambda} \sum_{i=1}^{n} (1 - f_i) \, f_i^2 \, \frac{\beta_i^2}{\sigma_i^2}, \qquad \rho' = \frac{4}{\lambda} \sum_{i=1}^{n} (1 - f_i)^2 \, f_i \, \beta_i^2 \tag{15}$$

where $\beta_i = u_i^T b$. These expressions follow from the relations

$$\frac{d\, f_i^2}{d\lambda} = -\frac{4}{\lambda} \, (1 - f_i) \, f_i^2 \qquad \text{and} \qquad \frac{d\, (1 - f_i)^2}{d\lambda} = \frac{4}{\lambda} \, (1 - f_i)^2 \, f_i.$$

Now using the fact that

$$\frac{f_i}{\sigma_i^2} = \frac{1}{\sigma_i^2 + \lambda^2} = \frac{1 - f_i}{\lambda^2}$$

we arrive at the important relation

$$\rho' = -\lambda^2 \, \eta'. \tag{16}$$

The next step involves the computation of the second derivatives of $\widehat{\eta}$ and $\widehat{\rho}$ with respect to λ, given by

$$\widehat{\eta}'' = \frac{d}{d\lambda} \frac{\eta'}{\eta} = \frac{\eta'' \eta - (\eta')^2}{\eta^2} \qquad \text{and} \qquad \widehat{\rho}'' = \frac{d}{d\lambda} \frac{\rho'}{\rho} = \frac{\rho'' \rho - (\rho')^2}{\rho^2}.$$

It suffices to consider the quantity

$$\rho'' = \frac{d}{d\lambda} \left(-\lambda^2 \, \eta' \right) = -2 \, \lambda \, \eta' - \lambda^2 \eta''. \tag{17}$$

When we insert all the expressions for $\widehat{\eta}'$, $\widehat{\eta}''$, $\widehat{\rho}'$, and $\widehat{\rho}''$ into the formula for κ and make use of (16) and (17), then both η'' and ρ'' as well as ρ' vanish, and we end up with the following expression

$$\kappa = 2\,\frac{\eta\,\rho}{\eta'}\,\frac{\lambda^2\,\eta'\,\rho + 2\,\lambda\,\eta\rho + \lambda^4\eta\,\eta'}{(\lambda^2\eta^2+\rho^2)^{3/2}}, \tag{18}$$

where the quantity η' is given by (15).

6 When the L-curve is concave

In Section 4 we made some arguments that an exact right-hand side $\overline{b}$ that satisfies the Discrete Picard Condition, or a right-hand side e corresponding to pure white noise, leads to an L-curve that is concave when plotted in log-log scale. In this section we present some new theory that supports this.

We restrict our analysis to an investigation of the circumstances in which the log-log L-curve is concave. Clearly, in any practical setting with a noisy right-hand side the L-curve cannot be guaranteed to be concave, and the key issue is in fact that the L-curve has an L-shaped (convex) corner for these right-hand sides. However, in order to understand the basic features of the L-curve it is still interesting to investigate its properties in connection with the idealized right-hand sides $b = \overline{b}$ and $b = e$.

Regińska made a first step towards such an analysis in [32] where she proved that the log-log L-curve is always strictly concave for $\lambda \le \sigma_n$ (the smallest singular value) and for $\lambda \ge \sigma_1$ (the largest singular value). Thus, the L-curve is always concave at its "ends" near the axes.

Our analysis extends Regińska's analysis. But instead of using $d^2\widehat{\eta}/d\widehat{\rho}^2$, we base our analysis on the above expression (18) for the curvature κ. In particular, if κ is negative then the L-curve is concave. Obviously, we can restrict our analysis to the factor $\lambda^2\,\eta'\,\rho + 2\,\lambda\,\eta\,\rho + \lambda^4\eta\,\eta'$, and if we define ζ such that

$$2\,\lambda\,\zeta = \lambda^2\,\eta'\,\rho + 2\,\lambda\,\eta\,\rho + \lambda^4\eta\,\eta',$$

and if we insert the definitions of η, η', and ρ into this expression, then we obtain

$$\zeta = \sum_{i=1}^{n}\sum_{j=1}^{n}(1-f_i)^2\,f_j^2\,\sigma_j^{-2}\;(2f_j - 2f_i - 1)\;\beta_i^2\,\beta_j^2, \tag{19}$$

where we have introduced $\beta_i = u_i^T b$.

Unfortunately we have not found a way to explicitly analyze Eq. (19). Our approach is therefore to replace the discrete variables σ_i, σ_j, β_i, and f_i with continuous variables σ, $\overline{\sigma}$, $\beta = \beta(\sigma)$, and $f = f(\sigma) = \sigma^2/(\sigma^2+\lambda^2)$, replace the double summation with a double integral, and study the quantity

$$\begin{aligned} \Xi \;=\; & \int_0^1 \int_0^1 \left(\frac{\lambda^2}{\sigma^2+\lambda^2}\right)^2 \left(\frac{\overline{\sigma}^2}{\overline{\sigma}^2+\lambda^2}\right)^2 \overline{\sigma}^{-2} \times \\ & \left(\frac{2\,\overline{\sigma}^2}{\overline{\sigma}^2+\lambda^2} - \frac{2\,\sigma^2}{\sigma^2+\lambda^2} - 1\right) \beta(\sigma)^2\, \beta(\overline{\sigma})^2 \, d\sigma \, d\overline{\sigma}. \end{aligned} \tag{20}$$

The sign of Ξ approximately determines the curvature of the L-curve. Without loss of generality, we assume that $\sigma_1 = 1$ and $0 \le \lambda \le 1$.

To simplify the analysis, we also make the assumption that β is simply given by

$$\beta(\sigma) = \sigma^{p+1}, \tag{21}$$

where p is a real number, and we denote the corresponding integral in (20) by Ξ_p. The model (21) is not unrealistic and models a wide range of applications. It is the basis of many studies of model problems in both the continuous and the discrete setting, see Section 4.5 in [21] for details. The quantity p controls the behavior of the right-hand side. The case $p > 0$ corresponds to a right-hand side that satisfies the Discrete Picard Condition (for example, an exact right-hand side $\overline{b}$), while $p \le 0$ corresponds to a right-hand side that does not satisfy the Discrete Picard Condition. In particular, $p = -1$ corresponds to a right-hand side e consisting of white noise. By means of Maple we can easily derive the following results.

Theorem 1 *Let Ξ_p and β be given by (20) and (21), respectively. Then for $p = -1$, $-1/2$, 0, $1/2$, and 1 we obtain:*

$$\begin{aligned} \Xi_{-1} &= -\frac{\lambda + \pi(1-\lambda^2)/4}{4\,\lambda^2\,(\lambda^2+1)^2} \\ \Xi_{-1/2} &= -\frac{1 + \lambda^2\,(2\ln\lambda - \ln(\lambda^2+1))}{4\,\lambda^2\,(\lambda^2+1)^2} \\ \Xi_0 &= -\frac{3\frac{\pi}{4}\lambda^2 - 3\lambda + \frac{\pi}{4}}{4\lambda\,(\lambda^2+1)^2} \\ \Xi_{1/2} &= -\frac{(2\lambda^2+1)\ln(\lambda^2+1) - 2\,(2\lambda^2-1)\ln\lambda - 2}{4\,(\lambda^2+1)^2} \end{aligned}$$

$$\Xi_1 \;=\; -\frac{4 - 15\frac{\pi}{4}\lambda^3 + 15\lambda^2 - 9\frac{\pi}{4}\lambda}{12\,(\lambda^2+1)^2}.$$

All these quantities are negative for $0 \leq \lambda \leq 1$.

The conclusion of this analysis is that as long as the SVD coefficients $|u_i^T b|$ decrease monotonically or increase monotonically with i, or are constant, then there is good reason to believe that the log-log L-curve is concave.

7 The L-curve criterion for computing the regularization parameter

The fact that the L-curve for a noisy right-hand side $b = \bar{b} + e$ has a more or less distinct corner prompted the author to propose a new strategy for choosing the regularization parameter λ, namely, such that the corresponding point on the L-curve

$$(\,\hat{\rho}/2\,,\;\hat{\eta}/2\,) = \Big(\log\|A\,x_\lambda - b\|_2\,,\;\log\|x_\lambda\|_2\Big)$$

lies on this corner [18]. The rationale behind this choice is that the corner separates the flat and vertical parts of the curve where the solution is dominated by regularization errors and perturbation errors, respectively. We note that this so-called *L-curve criterion* for choosing the regularization parameter is one of the few current methods that involve both the residual norm $\|A\,x_\lambda - b\|_2$ and the solution norm $\|x_\lambda\|_2$.

In order to provide a strict mathematical definition of the "corner" of the L-curve, Hansen and O'Leary [22] suggested using the point on the L-curve $(\hat{\rho}/2, \hat{\eta}/2)$ with maximum curvature κ given by Eq. (18). It is easy to use a one-dimensional minimization procedure to compute the maximum of κ. Various issues in locating the corner of L-curves associated with other methods than Tikhonov regularization are discussed in [22] and Section 7.5.2 in [21].

Figure 3 illustrates the L-curve criterion: the left part of the figure shows the L-curve, where the corner is clearly visible, and the right part shows the curvature κ of the L-curve as a function of λ. The sharp peak in the κ-curve corresponds, of course, to the sharp corner on the L-curve.

Experimental comparisons of the L-curve criterion with other methods for computing λ, most notably the method of generalized

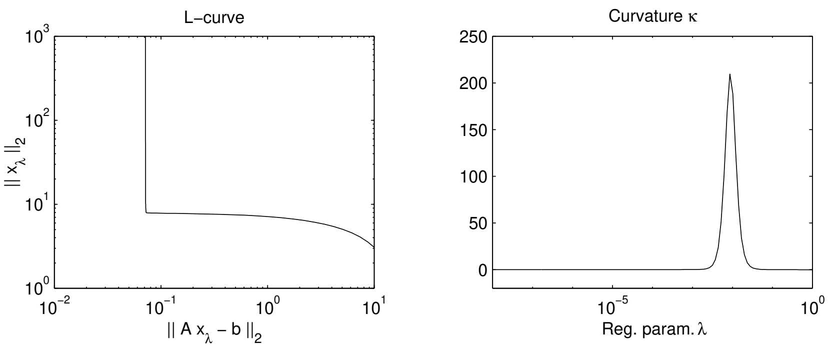

Figure 3: A typical L-curve (left) and a plot (right) of the corresponding curvature κ as a function of the regularization parameter.

cross validation (GCV) developed in [10] and [38], are presented in [22] and in Section 7.7.1 of [21]. The test problem in [22] is the problem shaw from the REGULARIZATION TOOLS package [19], [20], and the test problem in [21] is helio which is available via the author's home page. Both tests are based on ensembles with the same exact right-hand side $\bar{b}$ perturbed by randomly generated perturbations e that represent white noise.

The conclusion from these experiments is that the L-curve criterion for Tikhonov regularization gives a very robust estimation of the regularization parameter, while the GCV method occasionally fails to do so. On the other hand, when GCV works it usually gives a very good estimate of the optimal regularization parameter, while the L-curve criterion tends to produce a regularization parameter that slightly over-smooths, i.e., it is slightly too large.

Further experiments with correlated noise in [22] show that the L-curve criterion in this situation is superior to the GCV method which consistently produces severe under-smoothing.

The actual computation of κ, as a function of λ, depends on the size of the problem and, in turn, on the algorithm used to compute the Tikhonov solution x_λ. If the SVD of A can be computed then κ can readily be computed by means of Eqs. (8)–(9) and (15)–(18).

For larger problems the use of the SVD may be prohibitive while it is still feasible to compute x_λ via the least squares formulation (3). In this case we need an alternative way to compute the quantity η' in (15), and it is straightforward to show (by insertion of the SVD) that η' is given by

$$\eta' = \frac{4}{\lambda}\, x_\lambda^T z_\lambda, \qquad z_\lambda = \left(A^T A + \lambda^2 I\right)^{-1} A^T (A\, x_\lambda - b). \tag{22}$$

Hence, to compute η' we need the vector z_λ which is the solution to the problem

$$\min \left\| \begin{pmatrix} A \\ \lambda\, I \end{pmatrix} z - \begin{pmatrix} A\, x_\lambda - b \\ 0 \end{pmatrix} \right\|_2$$

and which can be computed by the same algorithm and the same software as x_λ. The vector z_λ is identical to the correction vector in the first step of iterated Tikhonov regularization, cf. Section 5.1.5 in [21].

For large-scale problems where any direct method for computing the Tikhonov solution is prohibitive, iterative algorithms based on Lanczos bidiagonalization are often used. For these algorithms the techniques presented in [1] and [12] can be used to compute envelopes in the $(\rho,\, \eta)$-plane that include the L-curve.

8 Limitations of the L-curve criterion

Every practical method has its advantages and disadvantages. The advantages of the L-curve criterion are robustness and ability to treat perturbations consisting of correlated noise. In this section we describe two disadvantages or limitations of the L-curve criterion; understanding these limitations is a key to the proper use of the L-curve criterion and, hopefully, also to future improvements of the method.

8.1 Smooth solutions

The first limitation to be discussed is concerned with the reconstruction of very smooth exact solutions, i.e., solutions $\overline{x}$ for which the corresponding SVD coefficients $|v_i^T \overline{x}|$ decay fast to zero, such that the solution $\overline{x}$ is dominated by the first few SVD components. For such solutions, Hanke [13] showed that the L-curve criterion will fail, and the smoother the solution (i.e., the faster the decay) the worse the λ computed by the L-curve criterion.

It is easy to illustrate and explain this limitation of the L-curve criterion by means of a numerical example. We shall use the test problem shaw from [19], [20] (see also Section 1.4.3 in [21]) with dimensions $m = n = 64$, and we consider two exact solutions: the solution $\overline{x}$

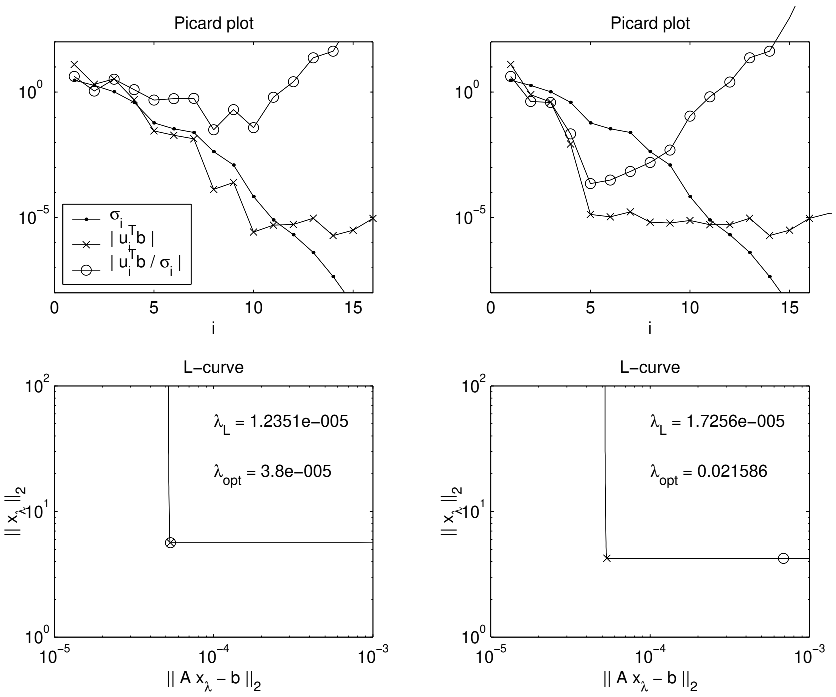

Figure 4: SVD coefficients and L-curves for a mildly smooth solution $\overline{x}$ (left) and a very smooth solution $\overline{\overline{x}}$ (right).

generated by **shaw**, which is mildly smooth, and a much smoother solution $\overline{\overline{x}} = \sigma_1^{-2} A^T A\,\overline{x}$ whose SVD coefficients are $v_i^T\overline{\overline{x}} = (\sigma_i/\sigma_1)^2 v_i^T\overline{x}$. The corresponding right-hand sides are $A\,\overline{x}+e$ and $A\,\overline{\overline{x}}+e$, where the elements of e are normally distributed with zero mean and standard deviation 10^{-5}. The two top plots in Fig. 4 show corresponding singular values and SVD coefficients for the two cases; note how much faster the SVD coefficients decay in the rightmost plot, before they hit the noise level.

The two bottom plots in Fig. 4 show the L-curves for the two cases. Located on the L-curves are two points: the corner as computed by means of the L-curve criterion (indicated by a $\times$ and corresponding to the regularization parameter λ_L) and the point corresponding to the optimal regularization parameter $\lambda_{\rm opt}$ (indicated by a $\circ$). Here, $\lambda_{\rm opt}$ is defined as the regularization parameter that minimizes the error $\|\overline{x} - x_\lambda\|_2$ or $\|\overline{\overline{x}} - x_\lambda\|_2$. For the problem with the mildly smooth solution $\overline{x}$ the L-curve criterion works well in the sense that $\lambda_L \simeq \lambda_{\rm opt}$, while for the problem with the smooth solution $\overline{\overline{x}}$ the regularization parameter λ_L is several orders of magnitude smaller than $\lambda_{\rm opt}$.

This behavior of the L-curve criterion is due to the fact that the optimal regularized solution $x_{\lambda_{\rm opt}}$ will only lie at the L-curve's corner if the norm $\|x_\lambda\|_2$ starts to increase as soon as λ becomes smaller than $\lambda_{\rm opt}$. Recall that λ controls roughly how many SVD components are included in the regularized solution x_λ via the filter factors f_i (6). If the exact solution is only mildly smooth, such as $\overline{x}$, and if the optimal regularized solution includes $\overline{k}$ SVD components, then only a few additional SVD components are required before $\|x_\lambda\|_2$ starts to increase. In Fig. 4 we see that $\lambda_L \simeq 1.2 \cdot 10^{-5}$ corresponds to including $\overline{k} = 10$ SVD components in x_λ, and decreasing λ by a factor of about 10 corresponds to including two additional large SVD components and thus increasing $\|x_\lambda\|_2$ dramatically. In addition we see that $\lambda_{\rm opt}$ also corresponds to including 10 SVD components in the optimal regularized solution, so $\lambda_{\rm opt}$ produces a solution $x_{\lambda_{\rm opt}}$ that lies at the corner of the L-curve.

If the exact solution is very smooth, such as $\overline{\overline{x}}$, and if the optimal regularized solution $x_{\lambda_{\rm opt}}$ includes $\overline{\overline{k}}$ SVD coefficients, then many additional coefficients may be required in order to increase the norm $\|x_\lambda\|_2$ significantly. The number of additional coefficients depends on the decay of the singular values. Returning to Fig. 4 we see that the optimal regularization parameter $\lambda_{\rm opt} \simeq 2.1 \cdot 10^{-2}$ for $\overline{\overline{x}}$ corresponds to including $\overline{\overline{k}} = 4$ SVD components in x_λ (the right-hand side's SVD components $u_i^T b$ are dominated by noise for $i > 4$). On the other hand, the regularization parameter $\lambda_L \simeq 1.7 \cdot 10^{-5}$ computed by the L-curve criterion corresponds to including 10 SVD components in x_λ, because at least 11 SVD components must be included before the norm $\|x_\lambda\|_2$ starts to increase significantly. Thus, the solution $x_{\lambda_{\rm opt}}$ does not correspond to a point on the L-curve's corner.

We note that for very smooth exact solutions the regularized solution $x_{\lambda_{\rm opt}}$ may not yield a residual whose norm $\|A\,x_\lambda - b\|_2$ is as small as $\mathcal{O}(\|e\|_2)$. This can be seen from the bottom right plot in Fig. 4 where $\|A\,x_{\lambda_{\rm opt}} - b\|_2 \simeq 6.9 \cdot 10^{-4}$ while $\|e\|_2 \simeq 6.2 \cdot 10^{-5}$, i.e., ten times smaller. The two solutions x_{λ_L} and $x_{\lambda_{\rm opt}}$ and their residuals are shown in Fig 5. Only x_{λ_L} produces a residual whose components are reasonably uncorrelated.

The importance of the quality of the fit, i.e., the size and the statistical behavior of the residual norm, depends on the application; but the dilemma between fit and reconstruction remains valid for very

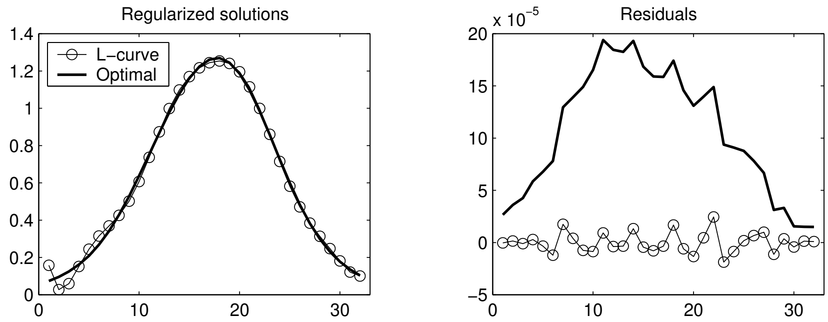

Figure 5: Test problem with a very smooth exact solution $\overline{\overline{x}}$: regularized solutions x_λ and corresponding residuals $b - A\,x_\lambda$ for λ_L computed by the L-curve criterion and $\lambda_{\rm opt}$ that minimizes $\|\overline{\overline{x}} - x_\lambda\|_2$.

smooth solutions. At the time of writing, it is not clear how often very smooth solutions arise in applications.

8.2 Asymptotic properties

The second limitation of the L-curve criterion is related to its asymptotic behavior as the problem size n increases. As pointed out by Vogel [37], the regularization parameter λ_L computed by the L-curve criterion may not behave consistently with the optimal parameter $\lambda_{\rm opt}$ as n increases.

The analysis of this situation is complicated by the fact that the linear system ($A\,x = b$ or $\min \|A\,x - b\|_2$) depends on the discretization method as well as the way the noise enters the problem. In our discussion below, we assume that the underlying, continuous problem is a first-kind Fredholm integral equation of the generic form

$$\int_0^1 K(s,t)\, f(t)\, dt = g(s), \qquad 0 \le s \le 1.$$

Here, the kernel K is a known function, the right-hand side g represents the measured signal, and f is the solution. We assume that $g = \overline{g} + \epsilon$, where $\overline{g}$ is the exact right-hand side and ϵ represents the errors. We also assume that the errors are white noise, i.e., uncorrelated and with the same standard deviation.

If the problem is discretized by a quadrature method then the elements x_i and b_i are essentially samples of the underlying functions

f and g, and the noise components e_i are samples of the noise ϵ. Then, as n increases, the quadrature weights included in A ensure that the singular values of A converge, while the SVD components $u_i^T b$ and $v_i^T x$ increase with n, and their magnitude is proportional to $n^{1/2}$. The sampled noise e consists of realizations of the same stochastic process and can be modeled by $u_i^T e = \epsilon$, where ϵ is a constant that is independent of n.

If the problem is discretized by means of a Galerkin method with orthonormal basis functions, then x_i and b_i are inner products of the chosen basis functions and the functions f and g, respectively. Then it is proved in [16] (see also Section 2.4 in [21]) that all the quantities σ_i, $u_i^T b$, and $v_i^T x$ converge as n increases. The noise components e_i can also be considered as inner products of basis functions and the noise component ϵ. If the noise is considered white, then we obtain the noise model $e_i = \epsilon$, where ϵ is a constant. If we assume that the norm of the errors ϵ is bounded (and then ϵ cannot be white noise) then we can use the noise model $e_i = \epsilon\, n^{-1/2}$, where again ϵ is a constant.

Vogel [37] considered a third scenario based on "moment discretization" in which σ_i increases as $n^{1/2}$ while $u_i^T b$ and $v_i^T x$ converge, and $e_i = \epsilon$. This is equivalent to the case $e_i = \epsilon\, n^{-1/2}$ immediately above.

To study these various scenarios in a common framework, we use the following simple model:

$$\left.\begin{array}{rcl} \sigma_i &=& \alpha^{i-1} \\ v_i^T \overline{x} &=& \beta^{i-1} \\ e_i &=& \epsilon\, n^{\gamma} \end{array}\right\} \qquad i = 1, \ldots, n$$

with $\alpha = 0.69,\ 0.83,\ 0.91$, $\beta = 0.95$, $\epsilon = 10^{-3}$, and $\gamma = 0,\ -1/2$. For $n = 100$ the three values of α yield a condition number of A equal to 10^{16}, 10^{8}, and 10^{4}, respectively, and β produces a mildly smooth solution. With $\gamma = 0$ the model represents white noise and discretization by means of a Galerkin method, and with $\gamma = -1/2$ the model represents the other scenarios introduced above.

For all six combinations of α and γ and for $n = 10^2$, 10^3, 10^4, and 10^5 we computed the optimal regularization parameter $\lambda_{\rm opt}$ as well as the parameter λ_L chosen by the L-curve criterion. The results are shown in Fig. 6, where the solid, dotted, and dashed lines correspond to $\alpha = 0.69$, 0.83, and 0.91, respectively. Recall that the smaller the α the faster the decay of the singular values.

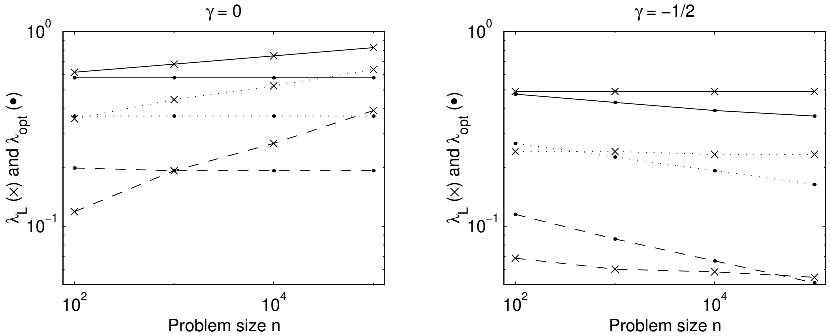

Figure 6: Plots of λ_L (crosses) and $\lambda_{\rm opt}$ (bullets) as functions of problem size n, for $\gamma = 0$ and $-1/2$ and for three values of α, namely, $\alpha = 0.69$ (solid lines), $\alpha = 0.83$ (dotted lines), and $\alpha = 0.91$ (dashed lines).

First of all note that the behavior of both λ_L and $\lambda_{\rm opt}$ depends on the noise model and the discretization method. For $\gamma = 0$ the parameter $\lambda_{\rm opt}$ is almost constant, while $\lambda_{\rm opt}$ decreases with n for $\gamma = -1/2$. The regularization parameter λ_L computed by the L-curve criterion increases with n for $\gamma = 0$, and for $\gamma = -1/2$ it is almost constant. Vogel [37] came to the same conclusion for the case $\gamma = -1/2$.

For all scenarios we see that the L-curve criterion eventually leads to some over-regularization (i.e., a too large regularization parameter) as n increases. However, the amount of over-smoothing depends on the decay of the singular values: the faster they decay the less severe the over-smoothing. Moreover, for $\gamma = -1/2$ and $n \leq 10^5$ the computed λ_L is never off the optimal value $\lambda_{\rm opt}$ by more than a factor 1.5 (while this factor is about two for $\gamma = 0$).

In conclusion, the over-smoothing that seems to be inherent in the L-curve criterion may not be too severe, although in the end this depends on the particular problem being solved. Another matter is that the ideal situation studied above, as well as in Vogel's paper [37], in which the *same* problem is discretized for increasing n, may not arise so often in practice. Often the problem size n is fixed by the particular measurement setup, and if a larger n is required then a new experiment must be made.

References

[1] D. Calvetti, G. H. Golub, and L. Reichel, Estimation of the L-curve via Lanczos bidiagonalization, *BIT*, 39 (1999), pp. 603–619.

[2] A. S. Carasso, Overcoming Hölder discontinuity in ill-posed continuation problems, *SIAM J. Numer. Anal.*, 31 (1994), pp. 1535–1557.

[3] L. Y. Chen, J. T. Chen, H.-K. Hong, and C. H. Chen, Application of Cesàro mean and the L-curve for the deconvolution problem, *Soil Dynamics and Earthquake Engineering*, 14 (1995), pp. 361–373.

[4] J. Cullum, The effective choice of the smoothing norm in regularization, *Math. Comp.*, 33 (1979), pp. 149–170.

[5] L. Eldén, Algorithms for the regularization of ill-conditioned least squares problems, *BIT*, 17 (1977), pp. 134–145.

[6] J. N. Franklin, Well-posed stochastic extensions of ill-posed linear problems, *J. Math. Anal. Appl.*, 31 (1970), pp. 682–716.

[7] J. N. Franklin, Minimum principles for ill-posed problems, *SIAM J. Math. Anal.*, 9 (1978), pp. 638–650.

[8] A. Frommer and P. Maass, Fast CG-based methods for Tikhonov-Phillips regularization, *SIAM J. Sci. Comput.*, 20 (1999), pp. 1831–1850.

[9] G. H. Golub, Numerical methods for solving linear least squares problems, *Numer. Math.*, 7 (1965), pp. 206–216.

[10] G. H. Golub, M. T. Heath, and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21 (1979), pp. 215–223.

[11] G. H. Golub and U. von Matt, Quadratically constrained least squares and quadratic problems, *Numer. Math.*, 59 (1991), pp. 561–579.

[12] G. H. Golub and U. von Matt, Tikhonov regularization for large scale problems; in G. H. Golub, S. H. Lui, F. T. Luk and R. J. Plemmons (Eds), *Scientific Computing*, Springer, Berlin, 1997; pp. 3–26.

[13] M. Hanke, Limitations of the L-curve method in ill-posed problems, *BIT*, 36 (1996), pp. 287–301.

[14] M. Hanke and C. R. Vogel, Two-level preconditioners for regularized inverse problems I: Theory, *Numer. Math.*, 83 (1999), pp. 385–402.

[15] P. C. Hansen, The truncated SVD as a method for regularization, *BIT*, 27 (1987), pp. 534–553.

[16] P. C. Hansen, Computation of the singular value expansion, *Computing*, 40 (1988), pp. 185–199.

[17] P. C. Hansen, The discrete Picard condition for discrete ill-posed problems, *BIT*, 30 (1990), pp. 658–672.

[18] P. C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Review*, 34 (1992), pp. 561–580.

[19] P. C. Hansen, Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems, *Numer. Algo.*, 6 (1994), pp. 1–35.

[20] P. C. Hansen, Regularization Tools version 3.0 for Matlab 5.2, *Numer. Algo.*, 20, (1999), pp. 195–196.

[21] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.

[22] 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. Comput.*, 14 (1993), pp. 1487–1503.

[23] A. E. Hoerl and R. W. Kennard, Ridge regression. Biased estimation for nonorthogonal problems, *Technometrics*, 12 (1970), pp. 55–67.

[24] A. E. Hoerl and R. W. Kennard, Ridge regression. Applications to nonorthogonal problems, *Technometrics*, 12 (1970), pp. 69–82.

[25] L. Kaufman and A. Neumaier, PET regularization by envelope guided conjugate gradients, *IEEE Trans. Medical Imaging*, 15 (1996), pp. 385–389.

[26] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentica-Hall, Englewood Cliffs, N.J., 1974; reprinted by SIAM, Philadelphia, 1995.

[27] D. W. Marquardt, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, 12 (1970), pp. 591–612.

[28] D. W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, *J. SIAM*, 11 (1963), pp. 431–441.

[29] K. Miller, Least squares methods for ill-posed problems with a prescribed bound, *SIAM J. Math. Anal.*, 1 (1970), pp. 52–74.

[30] A. Neumaier, Solving ill-conditioned and singular linear systems: A tutorial on regularization, *SIAM Review*, 40 (1998), pp. 636–666.

[31] D. L. Phillips, A technique for the numerical solution of certain integral equations of the first kind, *J. Assoc. Comput. Mach.*, 9 (1962), pp. 84–97.

[32] T. Regińska, A regularization parameter in discrete ill-posed problems, *SIAM J. Sci. Comput.*, 17 (1996), pp. 740–749.

[33] G. Ribiere, Regularisation d'operateurs, *R.I.R.O.*, 1 (1967), pp. 57–79.

[34] J. D. Riley, Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix, *Math. Tables Aids Comput.*, 9 (1955), pp. 96–101.

[35] A. N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Dokl.*, 4 (1963), pp. 1035–1038; English translation of *Dokl. Akad. Nauk. SSSR*, 151 (1963), pp. 501–504.

[36] S. Twomey, On the numerical solution of Fredholm integral equations of the first kind by inversion of the linear system produced by quadrature, *J. Assoc. Comput. Mach.*, 19 (1963), pp. 97–101.

[37] C. R. Vogel, Non-convergence of the L-curve regularization parameter selection method, *Inverse Problems*, 12 (1996), pp. 535–547.

[38] G. Wahba, Practical approximate solutions to linear operator equations when the data are noisy, *SIAM J. Numer. Anal.*, 14 (1977), pp. 651–667.