A REDUCED COMPUTATION METHOD FOR CHOOSING THE REGULARIZATION PARAMETER FOR TIKHONOV PROBLEMS

Soontorn Oraintara  William C. Karl  David A. Castanon  Truong Q. Nguyen
oraintar@bu.edu  wckarl@bu.edu  dac@bu.edu  nguyent@bu.edu

Electrical and Computer Engineering Department, Boston University
8 St. Mary’s St., Boston, MA 02215

ABSTRACT

This paper presents a method for choosing the regularization parameter (α) appearing in Tikhonov regularization based on the L-curve. The point of intersection between the L-curve and a straight line of arbitrary slope tangent to the L-curve is used as the criterion for identifying the corner of the L-curve and the corresponding value of α. This condition is shown to result in a new scalar algebraic equation for α in terms of the components of the SVD of the problem. A root of this equation yields the optimal α. The formulation of the problem in this way contrasts existing methods which require multiple solutions of the original regularization problem. Since computing the SVD remains costly for large problems, we then use the new method as the basis for rational approximation through the use of only a limited number of the singular values and vectors. Simulation results show that the proposed suboptimal method can provide a reasonable value of regularization parameter even when very few singular values and vectors are used in its definition.

1. INTRODUCTION

Linear inverse problems arise in many applications. Often, these problems can be represented by a Fredholm integral equation of the first kind:

\[ y(t) = \int c(t, \tau)x(\tau) d\tau + w(t) \tag{1} \]

where \( y(t) \) is the noisy observation, \( w(t) \) is additive noise, \( x(\tau) \) is the unknown object of interest, and \( c(t, \tau) \) is the kernel of the system. When discretized these problems lead to a vector-matrix representation of the form:

\[ y = Cx + w \tag{2} \]

Often, the kernel of the original problem is low-pass or smoothing, and the corresponding matrix \( C \) becomes ill-conditioned. In such cases, \( C \) is nearly singular, possessing small singular values, and the inverse problem is sometimes termed ill-posed.

A common way to solve (2) is to generate the estimated value of \( x \) as the least-square solution to the set of equations:

\[ x_{ls} = \arg \min_x ||y - Cx||^2, \tag{3} \]

where \( || \cdot || \) denotes the \( l_2 \) norm of a vector. When \( C \) is ill-conditioned, the least-square reconstructed object \( x_{ls} \) obtained through (3) will typically be corrupted by amplified noise and therefore unusable. In order to generate reasonable estimates some form of regularization is needed.

There are many methods to regularize such ill-conditioned problems. Among these methods, Tikhonov regularization is perhaps the most widely used, and defines the estimated object as the solution of an augmented minimization problem [1, 2, 3]:

\[ x_\alpha = \arg \min_x ||y - Cx||^2 + \alpha^2 ||Lx||^2 \tag{4} \]

The first term is just the least square data fit criterion as before. The second term is called the side constraint and captures prior knowledge about the unknown object, such as smoothness. The scalar parameter \( \alpha \) is called the regularization parameter and controls the tradeoff between fidelity to data and prior information. It is straightforward to show that the solution to (4) is given by:

\[ x_\alpha = (C^T C + \alpha^2 L^T L)^{-1} C^T y \triangleq R_\alpha y \tag{5} \]

where \( C^T \) is the transpose of \( C \).

A key element of any Tikhonov regularized formulation is the proper choice of the regularization parameter \( \alpha \). If \( \alpha \) is chosen to be too small (leading to what is termed under-regularization), the reconstruction will be too close to that obtained from Eq.(3) and thus dominated by large, high-frequency components. If \( \alpha \) is chosen to be too large (leading to what is termed over-regularization), the effect of the side-constraint term will dominate the solution and important information in the data will be suppressed.

While many approaches to choosing the regularization parameter have been presented in the literature, this paper focuses on methods based on what is called the L-curve [3, 4] - a log-log plot of the residual norm \( ||y - Cx_\alpha||^2 \) versus the regularization term \( ||Lx_\alpha||^2 \) as \( \alpha \) is varied. The name L-curve derives from the characteristic shape of this curve. In particular, when \( \alpha \) is very large (over-regularization), the residual norm is very sensitive to small changes in \( \alpha \) while the solution norm is relatively constant, so the curve is essentially a horizontal line. Conversely, when \( \alpha \) is very small (under-regularization), changes in the solution norm occur much faster than in the residual norm, and the curve is essentially a vertical line. Thus such a plot has a characteristic “L” shape. The transition between these two regions of
under and over-regularization corresponds to the “corner” of the L-curve and the associated value of $\alpha$ at this corner has been proposed as the optimal value of the regularization parameter. The L-curve approach is intuitive and has been shown to produce reasonable values of $\alpha$ in most situations.

Such an L-curve is shown in Figure 1. Practically then, this method of choosing the regularization parameter reduces to finding the corner of the L-curve. A number of different definitions have been proposed for this L-corner, including:

• The point of maximum curvature [5].
• The point closest to a reference point [6].
• The point of tangency with a straight line of negative slope [7].

Typically, computation of the corner of the L-curve requires repeated solution of the corresponding regularization problem for different values of $\alpha$, a potentially very costly task. In this paper, we outline an alternative formulation based on the last definition of the L-corner above. This formulation is new and obtains the L-corner as the root of a single scalar algebraic equation.

2. OPTIMUM REGULARIZATION PARAMETERS

Our method for identifying the optimal $\alpha$ is based on the last L-corner definition of the preceding section. That is, we define the corner as that point on the L-curve tangent to a line of angle $\theta$ with respect to the horizontal axis, as shown in Figure 1. The equation of a line $\ell$ in the L-curve coordinates can be described in terms of the parameters $\theta$ and $K$ by:

$$\log \|y - Cx\|^2 + \cot \theta \log \|Lx\|^2 = K$$

where $\theta$ is the angle of the line and $K$ is its intercept with the log-residual-norm axis. Given $\theta$ and $K$, the points on the intersection of this line and the L-curve are the solutions of the following equation:

$$\log \|y - CX_{\hat{\alpha}}\|^2 + \cot \theta \log \|LX_{\hat{\alpha}}\|^2 = K.$$ *(7)*

In general there will be a number of such intersection points. Our interest is in those near the L-corner. Let $x_{\hat{\alpha}}$ be the regularization parameter corresponding to the point of tangency of $\ell$ and the L-curve. Then $x_{\hat{\alpha}}$ can be obtained by minimizing $K$ subject to the constraint that $0 < \alpha < \infty$:

$$x_{\hat{\alpha}} = \arg \min_{0 < \alpha < \infty} K$$

$$= \arg \min_{0 < \alpha < \infty} \beta \log \|LX_{\hat{\alpha}}\|^2 + \log \|y - CX_{\hat{\alpha}}\|^2$$

$$= \arg \min_{0 < \alpha < \infty} \|y - CX_{\hat{\alpha}}\|^2 \|LX_{\hat{\alpha}}\|^2 \beta$$ *(8)*

where $\beta = \cot \theta$. The constraint on $\alpha$ is to assure that we obtain the local minimum corresponding to the corner of the L-curve.

Differentiating $K$ with respect to $\alpha$ and setting it to zero give a condition the optimum regularization parameter $x_{\hat{\alpha}}$ must satisfy:

$$\hat{\alpha}^2 \|LX_{\hat{\alpha}}\|^2 = \beta \|y - CX_{\hat{\alpha}}\|^2.$$ *(9)*

Figure 1: The Tikhonov L-curve illustrating the regularization parameter at the tangent point.

When $L = I$, this condition can be simplified with the aid of the SVD. The SVD of a matrix $C$ is a decomposition of $C$ of the form:

$$C = \sum_{i=1}^{m} \sigma_i u_i v_i^T$$ *(10)*

where the sets $\{u_i\}$ and $\{v_i\}$ are orthonormal and the $\sigma_i$ satisfy $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$. We have derived a scalar algebraic equation relating $\hat{\alpha}$ and the SVD components given by:

$$\hat{\alpha}^2 \sum_{i=1}^{m} (u_i^T y)^2 \frac{\beta \sigma_i^2 - \gamma_i^2}{(\hat{\alpha}^2 + \gamma_i^2)^2} = 0.$$ *(11)*

For the case when $L \neq I$, the generalized SVD (GSVD) can instead be used to simplify the problem. Suppose $C$ is $m \times n$, $L$ is $p \times n$ with $m \geq n \geq p$ and $L$ has full row rank. Then the GSVD [8] is a decomposition of $C$ and $L$ in the form:

$$C = U \times \Sigma \times V^T$$

$$L = U \times \Gamma$$ *(12)*

where $U_{m \times m}$ and $V_{p \times p}$ are unitary, and $\Sigma_{n \times n}$ is non-singular. $\Sigma = \text{diag}(\sigma_i)$ and $\Gamma = \text{diag}(\mu_i)$ are $p \times p$ diagonal matrices. Moreover

$$1 \geq \sigma_\ell \geq \cdots \geq \sigma_1 \geq 0,$$

$$0 \geq \mu_\ell \geq \cdots \geq \mu_1 \geq 0$$

with $\sigma_\ell^2 + \mu_\ell^2 = 1$. The values $\gamma_i = \frac{\alpha}{\sigma_i}$ are called the generalized singular values of $(C, L)$, and $\gamma_\ell \geq \cdots \geq \gamma_1 \geq 0$. Note that $\sigma_i$ and the columns $u_i$ and $v_j$ of $U$ and $V$ are different, in general, from the SVD.

In this case, a scalar algebraic equation relating $\hat{\alpha}$ and the GSVD components is given by:

$$\hat{\alpha}^2 \sum_{i=1}^{m} (u_i^T y)^2 \frac{\beta \gamma_i^2 - \gamma_i^2}{(\hat{\alpha}^2 + \gamma_i^2)^2} = -\sum_{i=n+1}^{m} (u_i^T y)^2.$$ *(13)*

The expressions Eq.(11) and Eq.(13) are new, simple scalar conditions for the optimal L-corner which explicitly
involve $\alpha$ and the singular components of the problem. Note that the resulting condition involves an equation of a scalar unknown, which can be solved numerically.

Finally, note that when $L = I$ the generalized singular values $\gamma_i$ reduce to the ordinary singular values $\sigma_i$ (except for ordering) and the generalized singular vectors reduce to the ordinary singular vectors (again, except for ordering), so that Eq.(13) reduces to Eq.(11) in this case.

2.1. Limiting the number of singular values and vectors

The conditions in (11) and (13) require the availability of the SVD (or GSVD) which is expensive to compute when $C$ is of large size, such as arises in the treatment of multi-dimensional problems. This computational need can be reduced by using only the first few (and therefore largest) singular values and corresponding singular vectors in the sums in (11) or (13). This use of only a few singular values essentially corresponds to truncating the sums in (11) or (13) after a few terms. Of course, truncating these sums will also affect the value of the corresponding estimated regularization parameter $\tilde{\alpha}$. As shown in the simulation, our experience is that we can obtain good estimates of $\tilde{\alpha}$ with quite few terms in (11) or (13).

2.2. Statistical view of the condition (9).

To provide some insight into the condition for the optimal value $\alpha$, let us provide a statistical interpretation of the equation (9). To this end, suppose that the underlying object $x$ is described by the prior model:

$$Lx = v, \quad v \sim N(0, \lambda_x I)$$

so that the process $Lx$ is zero-mean, white, and Gaussian with variance $\lambda_x$. Further suppose that the observation noise $w$ is specified by the following model:

$$w \sim N(0, \lambda_w I)$$

so that the noise is zero-mean, white, and Gaussian with variance $\lambda_w$. It is then straightforward to show that the maximum a-posteriori (MAP) estimate of $x$ based on the observation (2) and prior model (14) is given by:

$$\hat{x}_{\text{map}} = \arg \min_x ||y - Cx||^2 + \lambda_w \sigma_x ||Lx||^2.$$  

Notice that this expression is the same as obtained for Tikhonov regularization (4) if the regularization parameter is chosen as:

$$\alpha^2 = \frac{\lambda_w}{\lambda_x}.$$ 

That is, the optimal choice of the regularization parameter corresponding to this MAP problem is as the ratio of the noise standard deviation and prior model standard deviation.

Now the condition for the optimal regularization parameter in (9) when $\theta = 45^\circ$ (i.e. when we use a line of slope $-1$), is given by:

$$\alpha^2 = \frac{||y - C\hat{x}_\alpha||^2}{||Lx|\hat{\alpha}||^2}.$$  

If $x_\alpha$ well approximates $x$, then $||y - Cx_\alpha||^2$ can be viewed as an estimate of the noise variance and $||Lx_\alpha||^2$ can be viewed as an estimate of the prior variance. Thus the ratio in (18) appears consistent with the MAP choice of the regularization parameter above. Of course, the values appearing in (18) are the posterior quantities, and thus this interpretation is only reasonable in the case of little regularization.

3. GENERALIZATION TO MULTIPLE REGULARIZATION PARAMETERS

The previous development was for Tikhonov regularization with only a single regularization parameter. In many problems we are interested in introducing multiple constraints into the regularization formulation, which results in the need to specify multiple regularization parameters. For example, if the object is known to be a smooth signal, the Sobolev basis can be used to regularize the problem with different regularization parameter for each order of the derivative. Another example is frequency-domain analysis, e.g. Fourier or wavelet analysis where the vectors are projected into different subbands. Therefore it is natural generalization to consider the case of multiple-constraint Tikhonov problem. We now extend our previous results to the case of multiple regularization parameters. To this end, consider the following generalized Tikhonov regularization formulation:

$$x_\alpha = \arg \min_{x} ||y - Cx||^2 + \sum_{i=1}^{N} \alpha_i^2 ||L_i x||^2$$  

which is characterized by a vector of regularization parameters $\alpha = [\alpha_1, \ldots, \alpha_N]^T$. For a given set of $\alpha_i$, the optimum solution is given by:

$$x_\alpha = \left(C^T C + \sum_{i=1}^{N} \alpha_i^2 L_i^T L_i \right)^{-1} C^T y \overset{\Delta}{=} R_{\alpha} y.$$  

The notion of the L-curve can be naturally generalized to this multi-parameter situation by defining the hypersurface obtained as the log-scale plot of the residual norm $||y - Cx||^2$ versus the multiple constraint terms $||L_1 x||^2, \ldots, ||L_N x||^2$ as the $\alpha_i$ are varied. The “corner” of the resulting hypersurface can now be used to define the optimal choice of the regularization parameters.

Extending our previous approach to this multi-parameter case, we define the L-corner for this situation as the point of intersection of the hypersurface with the hyperplane:

$$\log ||y - Cx||^2 + \sum_{i=1}^{N} \beta_i \log ||L_i x||^2 = K$$

where $\beta_i > 0$ define the normal to the hyperplane and $K$ is again the intercept with the log-residual-norm axis. Proceeding as before, the optimum regularization parameters can be obtained as the solution of:

$$x_\alpha^* = \arg \min_{\alpha} ||y - Cx_\alpha||^2 \prod_{i=1}^{N} ||L_i x_\alpha||^{2\beta_i}.$$
In this example, we illustrate our development
In this example, a 2-D simulation is performed using the point of tangency methods.

The matrix \( C \) is obtained from this problem by simple quadrature discretization of the integral equation on a 120 \( \times \) 120 sampling lattice and the vectors \( x \) and \( y \) by corresponding uniform sampling on a uniform 120 point grid. The noise \( \psi \) in (2) is white Gaussian noise with signal to noise ratio (SNR) equals to 60 dB. The regularization matrix \( L \) is chosen to be the discrete first derivative matrix of size \( 119 \times 120 \). Figure 2 shows the L-corner of the problem.

In Figure 2, the L-corner determined by the maximum curvature criterion of [5] is plotted using the * mark. The corner corresponding to our proposed criteria with \( \theta = 45^\circ \) (\( \beta = 1 \)) is plotted as the × mark. The + mark corresponds to the reconstruction obtained by the GCV method, and the o mark is the position of the \( \alpha \) corresponding to minimum square error (MSE) reconstruction. The corresponding \( \alpha \) of the *, ×, o and + marks are 0.0214, 0.0255, 0.0115 and 0.0061 respectively. From Figure 2, we see that both L-curve criteria yield points close to the optimum one. However, to compute the location of the L-corner using the proposed approach, we only have to solve (13) for \( \alpha \) without generating the entire L-curve. The resulting reconstructions using various methods are presented in Figure 3.

Recall that, to be able to solve Eq.(13), the SVD (or GSVD) has to be computed, which is costly for problems of large size. We now examine the effect of truncing the summations in (13) by replacing \( p \) by \( k < p \). Figure 4 shows the value of \( \hat{\alpha} \) corresponding to different values of \( k \). It is evident that, as \( k \) increases, \( \hat{\alpha} \) quickly converges to its final optimal. In this example, the estimated \( \hat{\alpha} \) are essentially the same for \( 18 \leq k \leq p = 120 \).

Figure 5 shows the reconstructions corresponding to the values of \( \alpha_k \) found using different values of \( k \). Even using only a few singular values, a reconstruction close to that corresponding to the true L-corner can be obtained.

Example 2: In this example, a 2-D simulation is performed using the proposed method. Figure 6(a) show a 50 \( \times \) 50 original image. An isotropic Gaussian blur kernel is used in the experiment with 9-point support and 1.5 standard deviation. The noisy blurred image is shown in Figure 6(b) with SNR = 50 dB. The resulting reconstructions are presented in Figure 6(c)-(f) with the corresponding \( \hat{\alpha} \) and squared error summarized in Table 1. The terms MSE, GCV, MC and TG are respectively referred to the reconstruction using minimum squared error (exhausted searching), generalized cross validation, maximum curvature and the point of tangency methods.
Figure 3: The comparison the resulting reconstructions using various methods.

Figure 4: Approximation calculation of the regularization parameter $\alpha$ versus the number $k$ of singular values and vectors used in its calculation.

Table 1: Optimal regularization parameters and the corresponding squared error.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>MSE</th>
<th>GCCV</th>
<th>MC</th>
<th>TG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00398</td>
<td>238.905</td>
<td>239.311</td>
<td>254.290</td>
<td>239.897</td>
</tr>
</tbody>
</table>

5. CONCLUDING REMARKS

In this paper, we have presented a new method of finding the regularization parameter corresponding to the corner of the L-curve. We presented a new scalar algebraic equation characterizing this value of $\alpha$ in terms of the singular system of the underlying problem. This algebraic equation avoids the need for multiple solution evaluations typical of other methods. Further, we show that the algebraic equation may be approximated using only a few singular values and vectors, useful for large problems where the calculation of the entire singular system is very expensive. Experiments have shown that the resulting values of $\alpha$ and corresponding reconstructions are excellent approximations to those associated with the true L-corner.

6. REFERENCES

Figure 6: Reconstruction images using different methods in choosing the regularization parameter: (a) original, (b) noisy blurred image with 50 dB SNR, (c) best reconstruction (MSE), (d) using GCV, (e) using maximum curvature and (f) using the point of tangency.