3.6

Regularization in Image Restoration and Reconstruction

W. Clem Karl
Boston University

1 Introduction

This chapter focuses on the need for and use of regularization methods in the solution of image restoration and reconstruction problems. The methods discussed here are applicable to a variety of such problems. These applications to specific problems, including implementation specifics, are discussed in greater detail in the other chapters of the handbook. Our aim here is to provide a unifying view of the difficulties that arise, and the tools that are used, in the analysis and solution of these problems. In the remainder of this section a general model for common image restoration and reconstruction problems is presented together with the standard least-squares approach taken for solving these problems. A discussion of the issues leading to the need for regularization is provided. In Section 2 so-called direct regularization methods are treated while in Section 3 iterative methods of regularization are discussed. In Section 4 an overview is given of the important problem of parameter choice in regularization. Section 5 concludes the chapter.

1.1 Image Restoration and Reconstruction Problems

Image restoration and reconstruction problems have as their goal the recovery of a desired, unknown, image of interest $f(x, y)$ based on observation of a related set of distorted data $g(x, y)$. These problems are generally distinguished from image enhancement problems by their assumed knowledge and use of a *distortion model* relating the unknown $f(x, y)$ to the observed $g(x, y)$. In particular, we focus here on distortion models captured by a linear integral equation of the following form:

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y;x',y')f(x',y')\,dx'\,dy' \qquad (1)$$

where $h(x, y; x', y')$ is the kernel or response function of the distorting system, often termed the point spread function (PSF). Such a relationship is called a Fredholm integral equation of the first kind [1] and captures most situations of engineering interest. Note that equation (1), while linear, allows for the possibility of shift-variant system functions.

Examples of image *restoration* problems that can be captured by the distortion model (1) are discussed in Chapter 3.5 and include compensation for incorrect camera focus, removal of uniform motion blur, and correction of blur due to atmospheric turbulence. All these examples involve a spatially invariant PSF (that is, $h(x, y; x', y')$ is only a function of $x - x'$ and $y - y'$). One of the most famous examples of image restoration involving a *spatially varying* point spread function is provided by the Hubble Space Telescope, where a flaw in the main mirror resulted in a spatially varying distortion of the acquired images.

Examples of image *reconstruction* problems fitting into the framework of (1) include those involving reconstruction based on projection data. Many physical problems can be cast into this or a very similar tomographic-type framework, including: medical computer aided tomography, single photon emission tomography, atmospheric tomography, geophysical tomography, radio astronomy, and synthetic aperture radar imaging. The simplest model for these types of problems relates the observed projection $g(t, \theta)$ at angle θ and offset t to the underlying field $f(x, y)$ through a spatially-varying PSF given by [2]:

$$h(t, \theta; x', y') = \delta\big(t - x' \cos(\theta) - y' \sin(\theta)\big). \tag{2}$$

The set of projected data $g(t, \theta)$ is often called a sinogram. See Chapter 10.2 for more detail.

Even when the unknown field is modeled as continuous, the data, of necessity, are often discrete due to the nature of the sensor. Assuming there are N_g observations, (1) can be written as:

$$g_i = g(x_i, y_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_i(x', y')\, f(x', y')\, dx'\, dy' \quad 1 \le i \le N_g \tag{3}$$

where $h_i(x', y') = h(x_i, y_i; x', y')$ denotes the kernel corresponding to the i-th observation. In (3) the discrete data has been described as simple *samples* of the continuous observations. This can always be done, since any averaging induced by the response function of the instrument can be included in the specification of $h_i(x', y')$.

Finally, the unknown image $f(x, y)$ itself is commonly described in terms of a discrete and finite set of parameters. In particular, assume that the image can be adequately represented by a weighted sum of N_f basis functions $\phi_j(x, y)$, $j = 1, \ldots, N_f$ as follows:

$$f(x, y) = \sum_{j=1}^{N_f} f_j\, \phi_j(x, y). \tag{4}$$

For example, the basis functions $\phi_j(x, y)$ are commonly chosen to be the set of unit height boxes corresponding to an array of square pixels, though other bases (e.g. wavelets, see Chapter 4.2) have also found favor. Given the expansion in (4), the image is then represented by the collection of N_f coefficients f_j. For example, if a square $N \times M$ pixel array is used, then $N_f = NM$ and the f_j simply become the values of the pixels themselves.

Substituting (4) into (3) and simplifying yields the following completely discrete relationship between the set of observations g_i and the collection of unknown image coefficients f_j:

$$g_i = \sum_{j=1}^{N_f} H_{ij} f_j, \quad 1 \le i \le N_g \tag{5}$$

where H_{ij} is given by:

$$H_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_i(x', y')\, \phi_j(x', y')\, dx'\, dy', \quad 1 \le i \le N_g,\ 1 \le j \le N_f \tag{6}$$

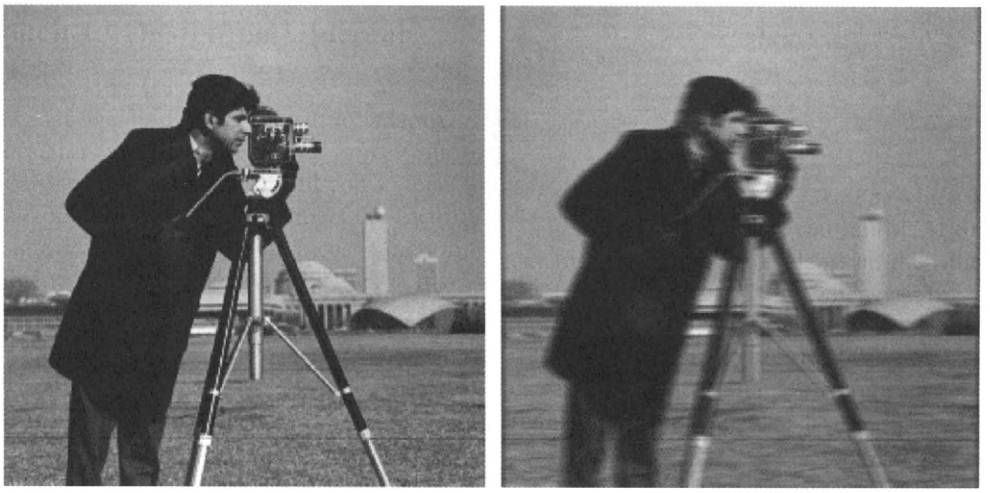
and represents the inner product of the i-th observation kernel $h_i(x, y)$ with the j-th basis function $\phi_j(x, y)$. Collecting all the observations g_i and image unknowns f_j into a single matrix equation yields a matrix equation capturing the observation process:

$$g = Hf \tag{7}$$

where the length N_g vector g, the length N_f vector f and the $N_g \times N_f$ matrix H follow naturally from (4), (5), and (6). When a rectangular pixel basis is used for $\phi_j(x, y)$ and the system is shift-invariant, so that $h(x, y; x', y') = h(x - x', y - y')$, the resulting matrix H will exhibit a banded block-Toeplitz structure with Toeplitz blocks – that is, the structure of H is of the form

$$H = \begin{bmatrix} H_0 & H_1 & H_2 & \cdots & H_M \\ H_{-1} & H_0 & H_1 & \cdots & H_{M-1} \\ H_{-2} & H_{-1} & H_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & H_1 \\ H_{-N} & \cdots & H_{-2} & H_{-1} & H_0 \end{bmatrix} \tag{8}$$

where the blocks H_i themselves internally exhibit the same banded Toeplitz structure. This structure is just a reflection of the linear convolution operation underlying the problem. Such linear convolutional problems can be represented by equivalent circular convolutional problems through appropriate zero padding. When this circulant embedding is done, the corresponding matrix H will then possess a block-circulant



(a) (b)

FIGURE 1 (a) Cameraman image (256 × 256). (b) Cameraman image distorted by 7-pixel horizontal motion blur and 30 dB SNR additive noise.

structure with circulant blocks [3]. In this case the structure of the associated H will be of the form:

$$H = \begin{bmatrix} H_0 & H_1 & H_2 & \cdots & H_M \\ H_M & H_0 & H_1 & \cdots & H_{M-1} \\ H_{M-1} & H_M & H_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & H_1 \\ H_1 & \cdots & H_{M-2} & H_{M-1} & H_0 \end{bmatrix} \quad (9)$$

where the block rows are simple circular shifts of each other and each block element itself possesses this structure. The significance of the block-circulant form is that there exist efficient computational approaches for problems with such structure in H, corresponding to the application of Fourier-based, frequency domain techniques.

In practice, our measured data is corrupted by inevitable perturbations or noise, so that our unknown image f is actually related to our data through:

$$g = Hf + q \quad (10)$$

where q is a vector of perturbation or noise samples. See Chapter 4.5 for models of image noise, including non-additive models. In what follows, the focus will be on the discrete or sampled data case as represented by (7) or (10).

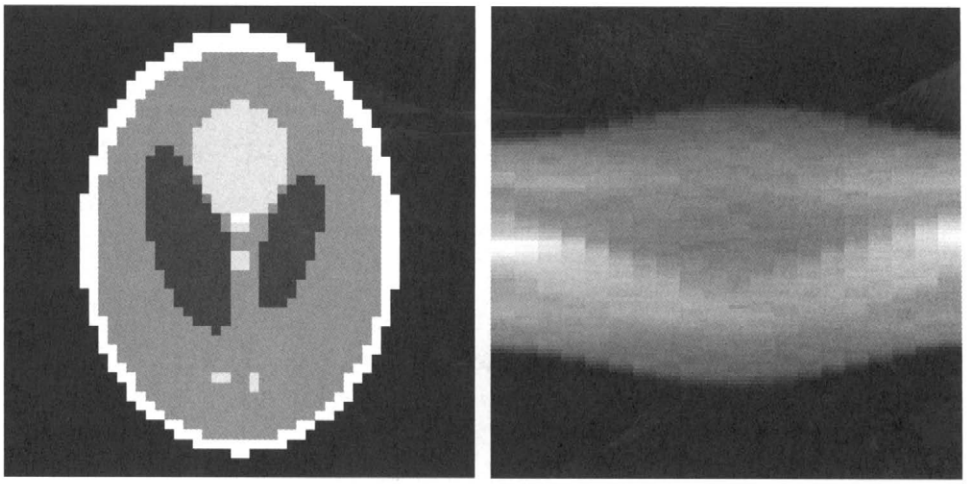
For purposes of illustration throughout the coming discussion, two example problems will be considered. The first example is an image restoration problem involving restoration of an image distorted by spatially invariant horizontal motion blur. The original 256 × 256 image is shown in Fig. 1(a). The distorted data, corresponding to application of a length 7-pixel horizontal motion blur followed by the addition of white Gaussian noise for an SNR[1] of 30 dB, is shown in Fig. 1(b).

The second example problem is a image reconstruction problem, involving reconstruction of an image from noisy tomographic data. The original 50 × 50 phantom image is shown in Fig. 2(a). The noisy projection data is shown in Fig. 2(b) with the horizontal axis corresponding to angle θ and the vertical axis corresponding to projection offset t. This data corresponds to application of (2) with 20 angles evenly spaced between $\theta = 0$ degrees and $\theta = 180$ degrees and 125 samples in t per angle followed by addition of white Gaussian noise for an SNR of 30 dB. This example represents a challenging inversion problem that might arise in non-destructive testing.

1.2 Least-Squares and Generalized Solutions

The image restoration or reconstruction problem can be seen to be equivalent to one of solving for the unknown vector f given knowledge of the data vector g and the distorting system matrix H. At first sight, a simple matrix inverse would seem to provide the solution, but this approach does not lead to usable solutions. There are four basic issues that must be dealt with in inverting the effects of H to find an estimate $\hat{f}$ of f. First, there is the problem of solution existence. Given an observation g in (7) there may not exist any f which solves this equation with equality, due to the presence of noise. Second, there is the problem of solution uniqueness. If the null-space of H is nonempty, then there are objects or images that are "unobservable" in the data. The null-space of H is the collection of all input images that produce zero output.

[1] SNR(dB) $\equiv 10 \log_{10}[\mathrm{Var}(Hf)/\mathrm{Var}(q)]$, where Var($z$) denotes the variance of z.



(a) (b)

FIGURE 2 (a) Original tomographic phantom image (50 × 50). (b) Projection data with 20 angles and 125 samples per angle corrupted by additive noise, 30 dB SNR.

An example would be the set of DC or constant images when H is a high-pass filter. Such components may exist in the true scene, but do not appear in the observations. In these cases, there will be many choices of f that produce the same set of observations and it must be decided which is the "right one." Such a situation arises, for example, when H represents a filter whose magnitude goes to zero for some range of frequencies, in which case images differing in these bands will produce identical observations. Third, there is the problem of solution stability. It is desired that the estimate of f remain relatively the same in the face of perturbations to the observations (either due to noise, uncertainty, or numeric roundoff). These first three elements are the basis of the classic Hadamard definition of an ill-posed problem [4, 5]. In addition to these problems, a final issue exists. Equation (7) only represents the observations, and says nothing about any prior knowledge about the solution. In general, more information will be available and a way to include it in the solution is needed. Regularization will prove to be the means to deal with all these problems.

The standard approach taken to inverting (7) will now be examined and its weaknesses explained in the context of the above discussion. The typical (and reasonable) solution to the first problem of solution existence is to seek a *least-squares solution* to the set of inconsistent equations represented by (7). That is, the estimate is defined as the least-squares fit to the observed data:

$$\hat{f}_{\text{ls}} = \arg\min_{f} \|g - Hf\|_2^2 \tag{11}$$

where $\|z\|_2^2 = \sum_i z_i^2$ denotes the ℓ_2-norm and arg denotes the argument producing the minimum (as opposed to the value of the minimum itself). A weighted error norm is also sometimes used in the specification of (11) to give certain observations increased importance in the solution: $\|g - Hf\|_W^2 = \sum_i w_i[g - Hf]_i^2$. If H has full column rank, the null-space of H is empty, and the estimate is unique and is obtained as the solution to the following set of normal equations [4]:

$$(H^T H)\hat{f}_{\text{ls}} = H^T g. \tag{12}$$

When the null-space of H is not empty, the second inversion difficulty of non-uniqueness, caused by the presence of unobservable images, must also be dealt with. What is typically done in these situations is to seek the unique solution of minimum energy or norm among the collection of least squares solutions. This *generalized solution* is usually denoted by $\hat{f}^+$ and defined as:

$$\hat{f}^+ = \arg\min_{f} \|f\|_2 \quad \text{subject to} \quad \min \|g - Hf\|_2. \tag{13}$$

The generalized solution is often expressed as $\hat{f}^+ = H^+ g$, where H^+ is called the generalized inverse of H (note that H^+ is defined implicitly through (13)). Thus generalized solutions are least-square solutions of minimum size or energy. Since components of the solution f that are unobservable do not improve the fit to the data, but only serve to increase the solution energy, the generalized solution corresponds to the least squares solution with no unobservable components, i.e., with no component in the null-space of H. Note that when the null-space of H is empty (for example, when we have at least as many independent observations g_i as unknowns f_j), the generalized and least-squares solutions are the same.

To understand how the generalized solution functions, consider a simple filtering situation where the underlying PSF

is shift-invariant (such as our deblurring problem) and the corresponding H is a circulant matrix. In this shift-invariant, filtering context the generalized solution method is sometimes referred to as "inverse filtering" (see Chapter 3.5). In this case, H can be diagonalized by the matrix F which performs the 2D discrete Fourier transform (DFT) on an image (represented as a vector) [3]. In particular, letting tildes denote transform quantities:

$$H = F^{-1}\widetilde{H}F, \quad H^T = F^{-1}\widetilde{H}^*F \tag{14}$$

where $\widetilde{H}$ is a diagonal matrix and $\widetilde{H}^*$ denotes the complex conjugate of $\widetilde{H}$. The diagonal elements of $\widetilde{H}$ are just the 2D DFT coefficients $\widetilde{h}_i$ of the PSF of this circulant problem (diag$[\widetilde{H}] = \widetilde{h} = Fh$, where h is given by, for example, the first column of H). Applying these relationships to (12), the following frequency domain characterization of the generalized solution is obtained:

$$\widetilde{H}^*\widetilde{H}\widetilde{f}^+ = \widetilde{H}^*\widetilde{g} \tag{15}$$

where $\widetilde{f}^+$ is a vector of the 2D DFT coefficients of the generalized solution, and $\widetilde{g}$ is a vector of the 2D DFT coefficients of the data. This set of equations is diagonal, so each component of the solution may be solved for separately:

$$\widetilde{f}_i^+ = \begin{cases} \left(\dfrac{1}{\widetilde{h}_i}\right)\widetilde{g}_i, & \left|\widetilde{h}_i\right| \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{16}$$

Thus, the generalized solution performs simple inverse filtering where the frequency response magnitude is non-zero, and sets the solution to zero otherwise.

For general, non-convolutional problems (for example, for tomographic problems) the 2D DFT matrix F does not provide a diagonalizing decomposition of H as in (14). There is, however, a generalization of this idea to arbitrary, shift-varying PSF system matrices called the singular value decomposition (SVD) [6]. The SVD is an important tool for understanding and analyzing inverse problems. The SVD of an $N_g \times N_f$ matrix H is a decomposition of the matrix H of the following form:

$$H = USV^T = \sum_{k=1}^{p} \sigma_k u_k v_k^T \tag{17}$$

where U is an $N_g \times N_g$ matrix, V is an $N_f \times N_f$ matrix, and S is an $N_g \times N_f$ diagonal matrix with the values $\sigma_1, \sigma_2, \ldots, \sigma_p$ arranged on its main diagonal and zeros elsewhere, where $p = \min(N_g, N_f)$. The orthonormal columns u_i of U are called the left singular vectors, the orthonormal columns v_i of V are called the right singular vectors, the σ_i are called the singular values, and the set of triples $\{\sigma_i, u_i, v_i\}$, $1 \leq i \leq p$ is called the singular system of H. Further, if r is the rank of H, then the σ_i satisfy:

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0. \tag{18}$$

The calculation of the entire SVD is too computationally expensive for general problems larger than modest size, though the insight it provides makes it a useful conceptual tool nonetheless. It is possible, however, to efficiently calculate the SVD for certain structured problems (such as for problems with a separable PSF [3]) or to calculate only parts of the SVD for general problems. Such calculations of the SVD can be done in a numerically robust fashion and many software tools exist for this purpose.

The SVD allows the development of an analysis similar to (16) for general problems. In particular, the generalized solution can be expressed in terms of the elements of the SVD as follows:

$$\hat{f}^+ = \sum_{i=1}^{r} \frac{u_i^T g}{\sigma_i} v_i. \tag{19}$$

This expression, valid for any H, whether convolutional or not, may be interpreted as follows. The observed data g is decomposed with respect to the set of basis images $\{u_i\}$ (yielding the coefficients $u_i^T g$). The coefficients of this representation are scaled by $1/\sigma_i$ and then used as the weights of an expansion of $\hat{f}^+$ with respect to the new set of basis images $\{v_i\}$. Note, in particular, that the sum only runs up to r. The components v_i of the reconstruction for $i > r$ correspond to $\sigma_i = 0$ and are omitted from the solution. These components correspond precisely to images that will be unobserved in the data. For example, if H were a low-pass filter, DC image components would be omitted from the solution. Note that for a linear shift-invariant problem, where frequency domain techniques are applicable, the solution (19) is equivalent to inverting the system frequency response at those frequencies where it is non-zero, and setting the solution to zero at those frequencies where it is zero, as previously discussed.

1.3 The Need for Regularization

A number of observations about the drawbacks of the generalized solution (13) or (19) may be made. First, the generalized solution makes no attempt to reconstruct components of the image that are unobservable in the data (i.e., in the null space of H). For example, if a given pixel is obscured from view, the generalized solution will set its value to zero despite the fact that all values near it might be visible (and hence, despite the fact that a good estimate as to its value may be

made). Second, and perhaps more seriously, the generalized solution is "unstable" in the face of perturbations to the data — that is, small changes in the data lead to large changes to the solution. To understand why this is so, note that most physical PSF system matrices H have the property that their singular values σ_i tend gradually to zero as i increases and, further, the singular image vectors u_i and v_i corresponding to these small σ_i are high-frequency in nature. The consequences of this behavior are substantial. In particular, the impact of the data on the i-th coefficient of the generalized solution (19) can be expressed as:

$$\frac{u_i^T g}{\sigma_i} = v_i^T f + \frac{u_i^T q}{\sigma_i}. \tag{20}$$

There are two terms on the right hand side of (20): the first term is due to the true image and the second term is due to the noise. For large values of the index i, these terms are like high frequency Fourier coefficients of the respective elements (since u_i and v_i are typically high frequency for large values of their index i). The high frequency contribution from the true image $v_i^T f$ will generally be much smaller than that due to the noise $u_i^T q$, since images tend to be lower frequency than noise. Further, the contribution from the noise is then *amplified* by the large factor $1/\sigma_i$. Overall then, the solution will be dominated by very large, oscillatory terms that are due to the noise. Another way of understanding this undesirable behavior follows from the generalized solution's insistence on reducing the data fit error above all else. If the data has noise, the solution $\hat{f}^+$ will be distorted in an attempt to fit to the noise components. Figure 3 shows the generalized solutions corresponding to the motion-blur restoration example of Fig. 1 and the tomographic reconstruction example of Fig. 2. The solutions have been truncated to the original range of the images in each case (either $[0, 255]$ for the motion-blur example or $[0, 1]$ for the tomographic example. Clearly these solutions are unacceptable.

The above insight provides not only a way of understanding why it is likely that the generalized inverse solution will have difficulties, but also a way of analyzing specific problems. In particular, since the generalized solution fails due to the explosion of the coefficients $u_i^T g/\sigma_i$ in the sum (19), potential inversion difficulties can be seen by plotting the quantities $|u_i^T g|$, σ_i and the ratio $|u_i^T g|/\sigma_i$ versus i [7]. Demonstrations of such plots for the two example problems are shown in Fig. 4. In both cases, for large values of the index i the coefficients $|u_i^T g|$ level off due to noise while the associated σ_i continue to decrease and thus the corresponding reconstruction coefficients in this range become very large.

It is for these reasons that the generalized solution is an unsatisfactory approach to the problems of image restoration and reconstruction in all but the lowest noise situations. These difficulties are generally a reflection of the ill-posed nature of the underlying continuous problem, as reflected in ill-conditioning of the system PSF matrix H. The answer to these difficulties is found through what is known as *regularization*. The purpose of regularization is to allow the inclusion of prior knowledge to stabilize the solution in the face of noise and allow the identification of physically meaningful and reasonable estimates. The basic idea is to constrain the solution in some way so as to avoid the oscillatory nature of the noise dominated solution observed in Fig. 3 [4].

A *regularization method* is often formally defined as an inversion method depending on a single real parameter $\alpha \geq 0$ which yields a family of approximate solutions $\hat{f}(\alpha)$ with the following two properties: first, for large enough α the regularized solution $\hat{f}(\alpha)$ is stable in the face of perturbations or noise in the data (unlike the generalized solution) and, second, as α goes to zero the unregularized generalized solution is recovered: $\hat{f}(\alpha) \longrightarrow \hat{f}^+$ as $\alpha \longrightarrow 0$. The parameter α is called the "regularization parameter" and controls the tradeoff between solution stability (i.e., noise propagation) and nearness of the regularized solution $\hat{f}(\alpha)$ to the unregularized solution $\hat{f}^+$ (i.e., approximation error in the absence of noise).

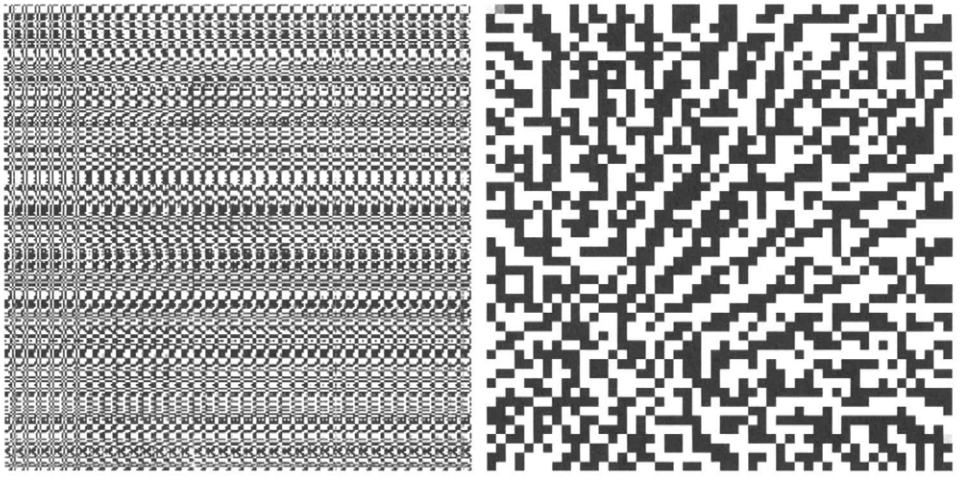


FIGURE 3 Generalized solutions $\hat{f}^+$ corresponding to data in Figs. 1(b) (left) and 2(b) (right).

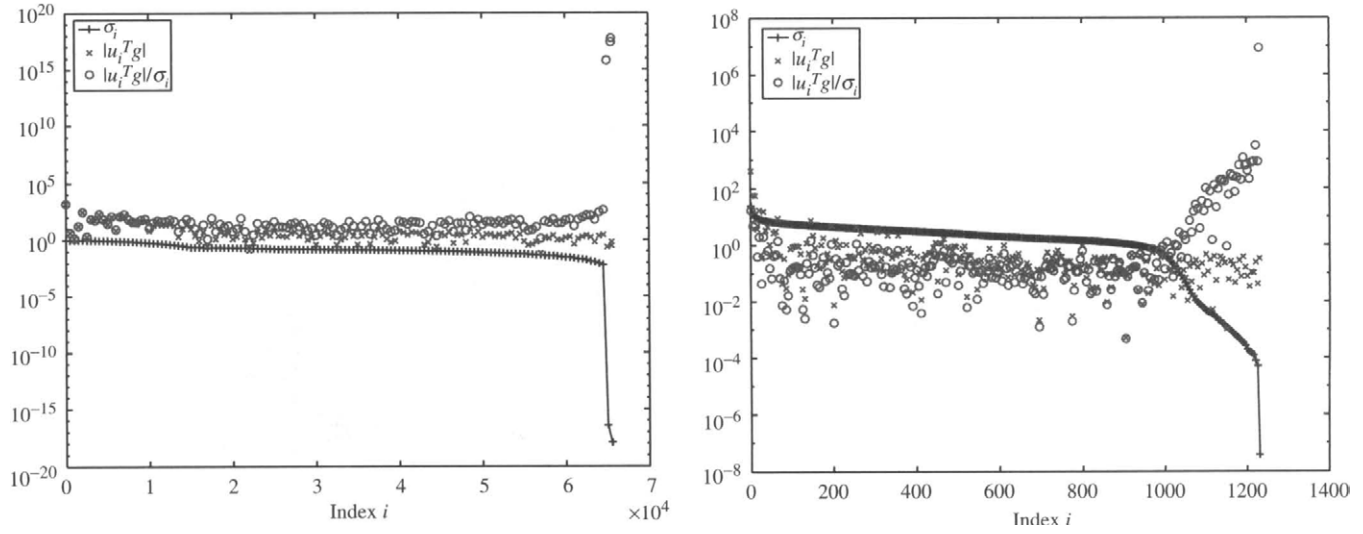


FIGURE 4 Plots of the components comprising the generalized solution for the problems in Figs. 1 (left) and 2 (right). (See color insert.)

Since the generalized solution represents the highest possible fidelity to the data, another way of viewing the role of α is in controlling the tradeoff between the impact of data and the impact of prior knowledge on the solution. There are a wide array of regularization methods, and an exhaustive treatment is beyond the scope of this chapter. The aim of this chapter is to provide a summary of the main approaches and ideas.

2 Direct Regularization Methods

In this section what are known as "direct" regularization methods are examined. These methods are conceptually defined by a direct computation, though they may utilize, for example, iterative methods in the computation of a practical solution. The iterative nature of any numeric solution is not intended to provide any additional regularizing effect, in this case. In Section 3 the use of iterative methods as a regularization approach in their own right is discussed.

2.1 Truncated SVD Regularization

From the discussion in Section 1.3, it can be seen that the stability problems of the generalized solution are associated with the large gain given the noise due to the smallest singular values σ_i. A logical remedy is to simply truncate small singular values to zero and thus remove the corresponding terms from the solution. This approach to regularization is called *truncated SVD* (TSVD) or numeric filtering [4, 7]. Indeed, such truncation is almost always done to some extent in the definition of the numeric rank of a problem, so TSVD simply does this to a greater extent. In fact, one interpretation of TSVD is as defining the rank of H *relative* to the noise in the problem. The TSVD regularized solution can be usefully defined based on (19) in the following way:

$$\hat{f}_{\text{tsvd}}(\alpha) = \sum_{i=1}^{r} w_{i,\alpha} \frac{u_i^T g}{\sigma_i} v_i \tag{21}$$

where $w_{i,\alpha}$ is a set of weights or filter factors given by:

$$w_{i,\alpha} = \begin{cases} 1 & i \le k(\alpha) \\ 0 & i > k(\alpha) \end{cases} \tag{22}$$

with the positive integer $k(\alpha) = \lfloor \alpha^{-1} \rfloor$, where $\lfloor x \rfloor$ denotes x rounded to the next smaller integer. Defined in this way, TSVD has the properties of a formal regularization method. TSVD simply throws the offending components of the solution out, but does not introduce any new components. As a result, TSVD solutions, while stabilized against noise, make no attempt to include image components that are unobservable in the data (like the original generalized solution).

Another way of understanding the TSVD solution is as follows. If H_k denotes the closest rank-k approximation to H, then, by analogy to (13), the TSVD solution $\hat{f}_{\text{tsvd}}(\alpha)$ in (19) is also given by:

$$\hat{f}_{\text{tsvd}}(\alpha) = \arg\min \|f\|_2 \quad \text{subject to} \quad \min \|g - H_k f\|_2, \tag{23}$$

which shows that the TSVD method can be thought of as directly approximating the original problem H by a nearby H_k which is better conditioned and less sensitive. In terms of its impact on reconstruction coefficients, the TSVD method corresponds to the choice of an ideal step weighting function $w_{i,\alpha}$ applied to the coefficients of the generalized solution. Certainly other weighting functions could and have been applied [4]. Indeed, some regularization methods are precisely interpretable in this way, as will be discussed.

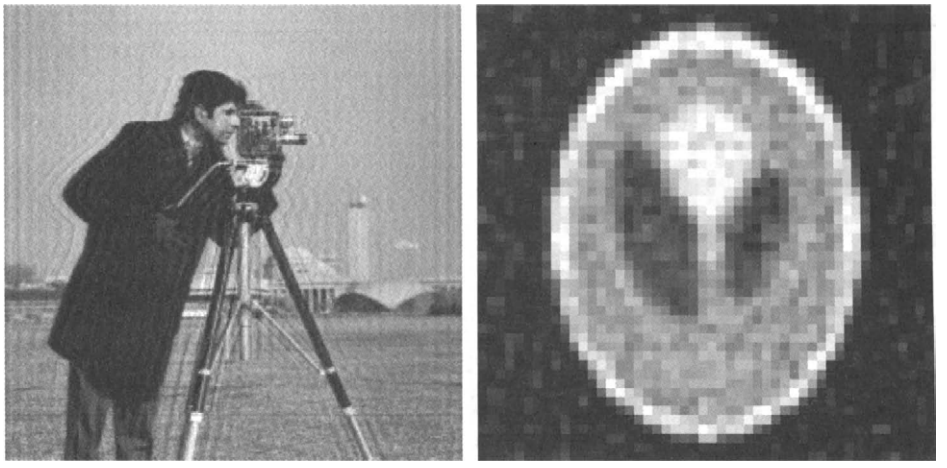


FIGURE 5 Truncated SVD solutions corresponding to data in Figs. 1(b) (left) and 2(b) (right).

Figure 5 shows truncated SVD solutions corresponding to the motion-blur restoration example of Fig. 1 and the tomographic reconstruction problem of Fig. 2. For the motion-blur restoration problem the solution used only approximately 40,000 of the over 65,000 singular values of the complete generalized reconstruction while for the tomographic reconstruction problem the solution used only about 800 of the full 2,500 singular values. As can be seen, the noise amplification of the generalized reconstruction has indeed been controlled in both cases. In the motion blur example some vertical ringing due to edge effects exists (see Chapter 3.5).

2.2 Tikhonov Regularization

Perhaps the most widely referenced regularization method is the Tikhonov method. The key idea behind the Tikhonov method is to directly incorporate prior information about the image f through the inclusion of an additional term to the original least-squares cost function. In particular, the Tikhonov regularized estimate is defined as the solution to the following minimization problem:

$$\hat{f}_{\text{tik}}(\alpha) = \arg\min_f \|g - Hf\|_2^2 + \alpha^2 \|Lf\|_2^2. \tag{24}$$

The first term in (24) is the same ℓ_2 residual norm appearing in the least-squares approach and ensures fidelity to data. The second term in (24) is called the "regularizer" or "side constraint" and captures prior knowledge about the expected behavior of f through an additional ℓ_2 penalty term involving just the image. The regularization parameter α controls the tradeoff between the two terms. The minimizer of (24) is the solution to the following set of normal equations:

$$(H^T H + \alpha^2 L^T L)\hat{f}_{\text{tik}}(\alpha) = H^T g. \tag{25}$$

This set of linear equations can be compared to the equivalent set (12) obtained for the unregularized least-squares solution. A solution to (25) exists and will be unique if the null spaces of H and L are distinct. There are a number of different numeric ways to obtain the Tikhonov solution from (25), including matrix inversion, iterative methods, and the use of factorizations like the SVD (or its generalizations) to diagonalize the system of equations.

To gain a deeper appreciation of the functioning of Tikhonov regularization, first consider the case when $L = I$, a diagonal matrix of ones. The corresponding side constraint term in (24) then simply measures the "size" or energy of f and thus, by inclusion in the overall cost function, directly prevents the pixel values of f from becoming too large (as happened in the unregularized generalized solution). The effect of α in this case is to trade off the fidelity to the data with the energy in the solution. Using the definition of the SVD combined with (25), the Tikhonov solution when $L = I$ can be expressed as:

$$\hat{f}_{\text{tik}}(\alpha) = \sum_{i=1}^{r} \left(\frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}\right) \frac{u_i^T g}{\sigma_i} \, v_i. \tag{26}$$

Comparing this expression to (19), an associated set of weight or filter factors $w_{i,\alpha}$ can be defined for Tikhonov regularization with $L = I$ as follows:

$$w_{i,\alpha} = \frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}. \tag{27}$$

In contrast to the ideal step behavior of the TSVD weights in (21), the Tikhonov weights decay like a "double-pole" low pass filter, where the "pole" occurs at $\sigma_i = \alpha$. Thus, Tikhonov regularization with $L = I$ can be seen to function similarly to TSVD, in that the impact of the higher index singular values on the solution is attenuated. Another consequence of this similarity is that when $L = I$, the Tikhonov solution again makes no attempt to reconstruct image components that are unobservable in the data.

The case when $L \neq I$ is more interesting.

Common choices for L include discrete approximations to the 2D gradient or Laplacian operators, resulting in measures of image slope and curvature, respectively. Such operators are discussed in Chapter 4.14. With these choices for L, $\|Lf\|$ is a measure of the "edginess" or roughness of the estimate. Inclusion of such terms in (24) forces solutions with limited high-frequency energy and thus captures a prior belief that solution images should be smooth. An expression for the Tikhonov solution when $L \neq I$ that is similar in spirit to (26) can be derived in terms of the generalized SVD of the pair (H, L) [6, 7], but is beyond the scope of this chapter. Interestingly, it can be shown that the Tikhonov solution when $L \neq I$ does contain image components that are unobservable in the data, and thus allows for extrapolation from the data. Note, it is also possible to consider the addition of multiple terms of the form $\|L_i f\|_2$, to create weighted derivative penalties of multiple orders, such as arise in Sobolev norms [4].

Figure 6 shows Tikhonov regularized solutions for both the motion-blur restoration example of Fig. 1 and the tomographic reconstruction example of Fig. 2 when $L = D$ is chosen as a discrete approximation of the gradient operator, so that the elements of Df are just the brightness changes in the image. The additional smoothing introduced through the use of a gradient-based L in the Tikhonov solutions can be seen in the reduced oscillation or variability of the reconstructed images.

Before leaving Tikhonov regularization it is worth noting that the following two inequality constrained least-squares problems are essentially the same as the Tikhonov method:

$$\hat{f} = \arg\min_f \|g - Hf\|_2 \text{ subject to } \|Lf\|_2 \leq 1/\lambda_1 \quad (28)$$

$$\hat{f} = \arg\min_f \|Lf\|_2 \text{ subject to } \|g - Hf\|_2 \leq \lambda_2. \quad (29)$$

The nonnegative scalars λ_1 and λ_2 play the roles of regularization parameters. The solution to each of these problems is the same as that obtained from (24) for a suitably chosen value of α that depends in a non-linear way on λ_1 or λ_2. The latter approach is also related to a method for choosing the regularization parameter called the "discrepancy principle," which we discuss in Section 4.

While (26) (and its generalization when $L \neq I$) gives an explicit expression for the Tikhonov solution in terms of the SVD, for large problems computation of the SVD may not be practical and other means must be sought to solve (25). When H and L have circulant structure (corresponding to a shift-invariant filter), these equations are diagonalized by the DFT matrix and the problem can be easily solved in the frequency domain. Often, even when this is not the case, the set of equations (25) possess a sparse and banded structure and may be efficiently solved using iterative schemes, such as preconditioned conjugate gradient.

2.3 Non-Quadratic Regularization

The basic Tikhonov method is based on the addition of a quadratic penalty $\|Lf\|_2$ to the standard least-squares (and hence quadratic) data fidelity criterion, as shown in (24). The motivation for this addition was the stabilization of the generalized solution through the inclusion of prior knowledge in the form of a side constraint. The use of such quadratic, ℓ_2-based criteria for the data and regularizer leads to the linear problem (25) for the Tikhonov solution, and thus results in an inverse filter which is a linear function of the data. While such linear processing is desirable, since it leads to straightforward and reasonably efficient computation methods, it is also limiting, in that far more powerful results are possible if non-linear methods are allowed. In particular, when used for suppressing the effect of high-frequency noise, such linear filters, by their nature, also reduce high frequency energy in the true image and hence blur detail in the reconstruction. For this reason, the generalization of the Tikhonov approach through the inclusion of certain non-quadratic criteria is now

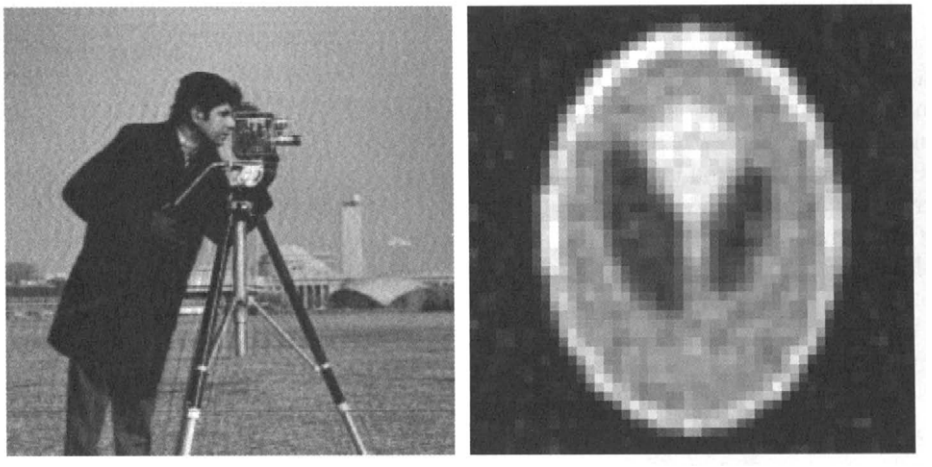


FIGURE 6 Tikhonov regularized solutions when L is a gradient operator corresponding to the data in Figs. 1(b) (left) and 2(b) (right).

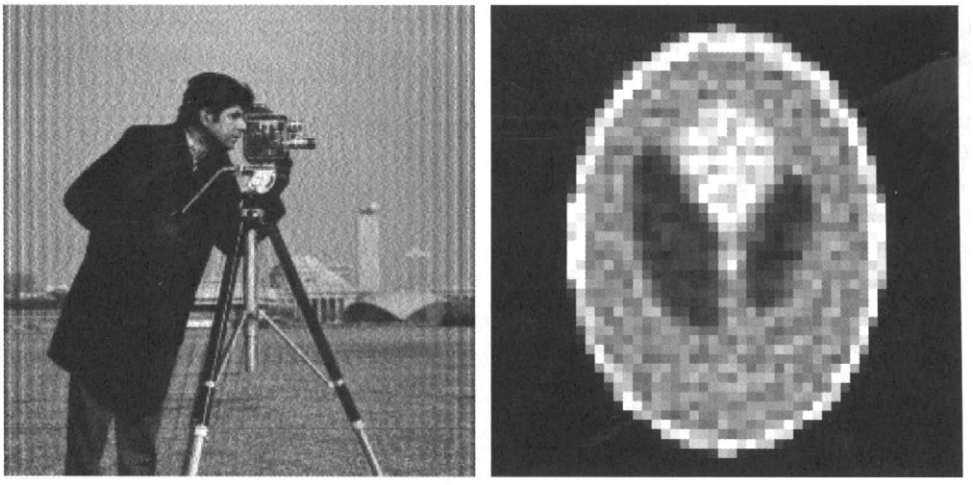


FIGURE 7 Maximum entropy solutions corresponding to the data in Figs. 1(b) (left) and 2(b) (right).

considered. To this end, consider estimates obtained as the solution of the following generalized formulation:

$$\hat{f}(\alpha) = \arg\min_f J_1(f, g) + \alpha^2 J_2(f) \tag{30}$$

where $J_1(f, g)$ represents a general distance measure between the data and its prediction based on the estimated f and $J_2(f)$ is a general regularizing penalty. Both costs may be a non-quadratic function of the elements of f. Next, a number of popular and interesting choices for $J_1(f, g)$ and $J_2(f)$ are examined.

Maximum Entropy Regularization

Perhaps the most widely used non-quadratic regularization approach is the maximum entropy method. The entropy of a positive valued image in the discrete case may be defined as:

$$-J_2(f) = -\sum_{i=1}^{N_f} f_i \log(f_i) \tag{31}$$

and can be taken as a measure of the uncertainty in the image. This interpretation follows from information theoretic considerations when the image is normalized so that $\sum_{i=1}^{N_f} f_i = 1$, and may thus be interpreted as a probability density function [5]. In this case, it can be argued that the maximum entropy solution is the most noncommittal with respect to missing information. A simpler motivation for the use of the entropy criterion is that it assures positive solutions. Combining the entropy cost (31) with a standard quadratic data fidelity term for $J_1(f, g)$ yields the maximum entropy estimate as the solution of:

$$\hat{f}_{\text{me}}(\alpha) = \arg\min_f \|g - Hf\|_2^2 + \alpha^2 \sum_{i=1}^{N_f} f_i \log(f_i). \tag{32}$$

There are a number of variants on this idea involving related definitions of entropy, cross-entropy, and divergence [5]. Experience has shown that this method provides image reconstructions with greater energy concentration (i.e., most coefficients are small and a few are very large) relative to quadratic Tikhonov approaches. For example, when the f_i represent pixel values, the approach has resulted in sharper reconstructions of point objects, such as star fields in astronomical images. The difficulty with the formulation (32) is that it leads to a nonlinear optimization problem for the solution, which must be solved iteratively.

Figure 7 shows maximum entropy solutions corresponding to both the motion-blur restoration example of Fig. 1 and the tomographic reconstruction example of Fig. 2. Note that these two examples are not particularly well matched to the maximum entropy approach, since in both cases the true image is not composed of point-like objects. Still, the maximum entropy side constraint has again succeeded in controlling the noise amplification observed in the generalized reconstruction. For the tomography example, note that small variations in the large background region have been supressed and the energy in the reconstruction has been concentrated within the reconstructed object. In the motion blur example, the central portion of the reconstruction is sharp, but the edges again show some vertical ringing due to boundary effects.

Total Variation Regularization

Another non-quadratic side constraint that has achieved popularity in recent years is the total variation measure:

$$J_2(f) = \|Df\|_1 = \sum_{i=1}^{N_f} \left| [Df]_i \right| \tag{33}$$

where $\|z\|_1$ denotes the ℓ_1-norm (i.e., the sum of the absolute values of the elements), D is a discrete approximation to the gradient operator described in Chapter 4.11, so that the

elements of Df are just the brightness changes in the image. The total variation of a signal is thus just the total amount of change the signal goes through and can be thought of as a measure of signal variability. Thus, it is well suited to use as a side constraint.

The corresponding total variation estimate is obtained by combining (33) with the standard quadratic data fidelity term for $J_1(f,g)$ to yield:

$$\hat{f}_{\text{tv}}(\alpha) = \arg\min_f \|g - Hf\|_2^2 + \alpha^2 \sum_{i=1}^{N_f} \left|[Df]_i\right|. \tag{34}$$

The formulation seems similar to standard Tikhonov regularization with a derivative constraint. But, unlike standard quadratic Tikhonov solutions, total variation regularized answers can contain localized steep gradients, since the resularizer penalizes only the total amount of gradient in the image and not its distribution. As a result, edges are preserved in the reconstructions. For these reasons, total variation has been suggested as the "right" regularizer for image reconstruction problems [8].

The difficulty with the formulation (34) is that it again leads to a challenging non-linear optimization problem due to the non-differentiability of the total variation cost. One approach to overcoming this challenge leads to an interesting formulation of the total variation problem. It has been shown that the total variation estimate is the solution of the following set of equations in the limit as $\beta \longrightarrow 0$:

$$\left(H^T H + \alpha^2 D^T W_\beta(\hat{f}_{\text{tv}}) D\right)\hat{f}_{\text{tv}} = H^T g \tag{35}$$

where the diagonal weight matrix $W_\beta(f)$ depends on f and β and is given by:

$$W_\beta(f) = \frac{1}{2}\text{diag}\left[\frac{1}{\sqrt{|[Df]_i|^2 + \beta}}\right] \tag{36}$$

with $\beta > 0$ a constant. Equation (35) is obtained by smoothly approximating the ℓ_1 norm of the derivative: $\|Df\|_1 \approx \sum_{i=1}^{n} \sqrt{|[Df]_i|^2 + \beta}$.

The formulation (36) is interesting in that it gives insight into the difference between total variation regularization and standard quadratic Tikhonov regularization with $L = D$. Note that the latter case would result in a set of equations similar to (35) but with $W = I$. Thus, the effect of the change to a total variation cost is the incorporation of a spatially varying weighting of each derivative penalty term by $1/\sqrt{|[Df]_i|^2 + \beta}$. When the local derivative $|[Df]_i|^2$ is small, the weight goes to a large value, imposing greater smoothness to the solution in these regions. When the local derivative $|[Df]_i|^2$ is large, the weight goes to a small value, allowing large gradients in the solution coefficients at these points.

Computationally, the equation (35) is still non-linear, since the weight matrix depends on f. However, it suggests a simple fixed point iteration for f, only requiring the solution of a standard linear problem at each step:

$$\left(H^T H + \alpha^2 D^T W_\beta(f^{(k)}) D\right) f^{(k+1)} = H^T g \tag{37}$$

where β is typically set to a small value.

Using the iterative approach of (37), total variation solutions to the motion-blur restoration example of Fig. 1 and the tomographic reconstruction example of Fig. 2 were generated. Figure 8 shows these total variation solutions. In addition to suppressing excessive noise growth, total variation achieves excellent edge preservation and structure recovery in both cases. These results are typical of total variation reconstructions, and has lead to the great popularity of this and similar methods in recent years.

Other Non-Quadratic Regularization

More generally, a variety of non-quadratic choices for $J_1(f,g)$ and $J_2(f)$ have been considered. In general, these measures have the characteristic that they do not penalize large values

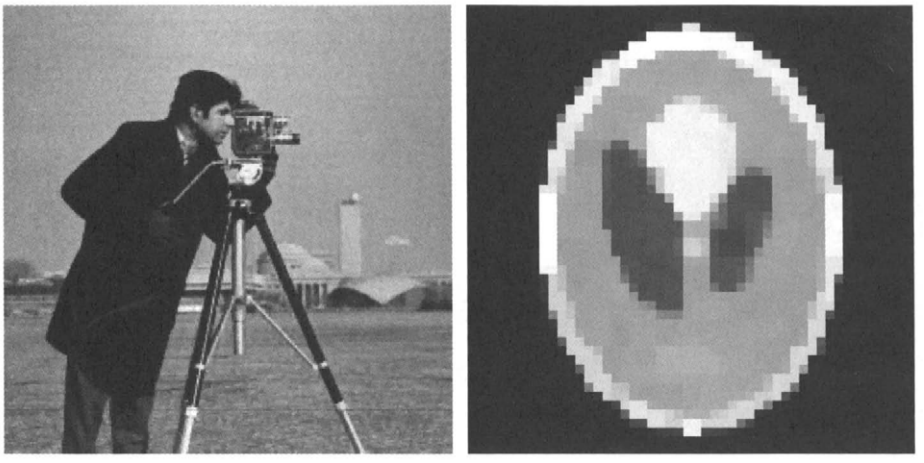


FIGURE 8 Total variation solutions corresponding to the data in Figs. 1(b) (left) and 2(b) (right).

of their argument as much as the standard quadratic ℓ_2 penalty does. Indeed, maximum entropy and total variation can both be viewed in this context, in that both are simple changes of the side constraint to a size or energy measure that is less drastic than squared energy. It has also been shown that the effect of these regularizers is that they lead to *sparsity* (also termed "near-blackness") of their argument [9]. For example, if applied as $J(Lf)$, the reconstruction will exhibit a sparse structure in Lf. That is, Lf will tend to have few large components and many near-zero elements. The choice of L determines the specific effect. If $L=I$ then the reconstruction will exhibit point-like structure. If $L=D$ then the reconstructed image will have sparse edges and sizable uniform (i.e., $Df \approx 0$) regions. We discuss some penalties with these properties next.

A regularizer choice with these properties is the general family of ℓ_p-norms:

$$J(z) = \|z\|_p^p = \sum_{i=1}^{N_z} |z_i|^p \tag{38}$$

with $1 \le p \le 2$. With p chosen in this range these norms are less severe in penalizing large values than the ℓ_2 norm, yet are still convex functions of the argument (and so still result in straight-forward algorithms). Note that Total Variation regularization is just the use of an ℓ_1 norm applied to the quantity Df.

Measures with even more drastic penalty "attenuation," based on non-convex functions, have been considered. One example is the so called "weak-membrane" cost:

$$J(z) = \min\left(\|z\|_2^2, \alpha\right). \tag{39}$$

Other examples, discussed in [10], include:

$$J(z) = \sum_{i=1}^{N_z} \frac{z_i^2}{1+z_i^2} \tag{40}$$

$$J(z) = \sum_{i=1}^{N_z} \log(1+z_i^2). \tag{41}$$

When used in the data fidelity term $J_1(f,g)$ these non-convex measures are related to notions of robust estimation [11] and provide robustness to outliers in the data and also to model uncertainty. When used in a side constraint $J_2(f)$ on the gradient of the image Df, these measures preserve edges in regularized reconstructions and produce results similar in quality to the total variation solution discussed previously. The difficulty with the use of such non-convex costs is computational, though the search for efficient approaches to such problems has been the subject of active study [12].

2.4 Statistical Methods

We now discuss a statistical view of regularization. If the noise q and the unknown image f are viewed as random fields, then we may seek the maximum a posteriori (MAP) estimate of f as that value which maximizes the posterior density $p(f \mid g)$. Using Bayes rule and the monotonicity properties of the logarithm we obtain:

$$\hat{f}_{\mathsf{MAP}} = \arg\max_f p(f \mid g) = \arg\max_f \ln p(g \mid f) + \ln p(f). \tag{42}$$

Notice that this cost function has two terms: a data dependent term $\ln p(g \mid f)$ called the log-likelihood function and a term $\ln p(f)$ dependent only on f termed the prior model. These two terms are similar to the two terms in the Tikhonov functional (24). The likelihood function captures the dependence of the data on the field, and enforces fidelity to data in (42). The prior model term captures our a priori knowledge about f in the absence of data, and allows incorporation of this information into the estimate.

To be concrete, consider the widely used case of Gaussian statistics:

$$g = Hf + q, \quad q \sim N(0, \Lambda_q) \tag{43}$$

$$f \sim N(0, \Lambda_f) \tag{44}$$

where $f \sim N(m, \Lambda)$ denotes that f is a Gaussian random vector with mean m and covariance matrix Λ, as described in Chapter 4.3. Under these assumptions $\ln p(g \mid f) \propto -\frac{1}{2}\|g - Hf\|^2_{\Lambda_q^{-1}}$ and $\ln p(f) \propto -\frac{1}{2}\|f\|^2_{\Lambda_f^{-1}}$ and upon substitution into (42) we obtain:

$$\hat{f}_{\mathsf{MAP}} = \arg\min_f \|g - Hf\|^2_{\Lambda_q^{-1}} + \|f\|^2_{\Lambda_f^{-1}} \tag{45}$$

The corresponding set of normal equations defining the MAP estimate are given by:

$$\left(H^T\Lambda_q^{-1}H + \Lambda_f^{-1}\right)\hat{f}_{\mathsf{MAP}} = H^T\Lambda_q^{-1}g \tag{46}$$

The solution of (46) is also the linear minimum mean-square error (MMSE) estimate in the general (i.e., non-Gaussian) case. The MMSE estimate minimizes $E[\|f - \hat{f}\|_2^2]$, where $E[z]$ denotes the expected value of z.

A particularly interesting prior model for f corresponds to:

$$Df = r, \quad r \sim N(0, \lambda_r I) \tag{47}$$

where D is a discrete approximation of the gradient operator described in Chapter 4.14. Equation (47) implies a Gaussian prior model for f with covariance $\Lambda_f = \lambda_r\left(D^TD\right)^{-1}$

(assuming, for convenience, that D is invertible). The prior model (47) essentially says that the *increments* of f are uncorrelated with variance λ_r — that is, that f itself corresponds to a Brownian motion-type model. Clearly, real images are not Brownian motions, yet the continuity of Brownian motion models suggest this model may be reasonable for image restoration.

To demonstrate this insight, the Brownian motion prior image model of (47) is combined with an uncorrelated observation noise model $\Lambda_q = \lambda_q I$ in (43) to obtain a MAP estimate for both the motion-blur restoration example of Fig. 1 and the tomographic reconstruction example of Fig. 2. In each case, the variance λ_r is set to the variance of the derivative image Df and the variance λ_q is set to the additive noise variance. In Fig. 9 the resulting MAP-based solutions for the two examples are shown.

In addition to an estimate, the statistical framework also provides an expression for an associated measure of estimate uncertainty through the error covariance matrix $\Lambda_e = E[ee^T]$, where $e = f - \hat{f}$. For the MAP estimate in (45), the error covariance is given by:

$$\Lambda_e = \left(H^T \Lambda_q^{-1} H + \Lambda_f^{-1}\right)^{-1}. \tag{48}$$

The diagonal entries of Λ_e are the variances of the individual estimation errors, and have a natural use for estimate evaluation and data fusion. Note also that the trace of Λ_e is the mean square error of the estimate. In practice, Λ_e is usually a large, full matrix, and the calculation of all its elements is impractical. There are methods, however, to estimate its diagonal elements [13].

In the stationary case, the matrices in (46) possess a block circulant structure and the entire set of equations can be solved on an element-by-element basis in the discrete frequency domain, as was done c.f. (16). The MAP estimator then reduces to the Wiener filter:

$$\tilde{f}_i = \left(\frac{\tilde{h}_i^* S_{f_i}}{\left|\tilde{h}_i\right|^2 S_{f_i} + S_{q_i}}\right) \tilde{g}_i \tag{49}$$

where $*$ denotes complex conjugate, $\tilde{z}_i$ denotes the i-th coefficient of the 2D DFT of the corresponding image z, and S_{z_i} denotes the i-th element of the power spectral density of the random image z. The power spectral density is the 2-D DFT of the corresponding covariance matrix. The Wiener filter is discussed in Chapter 3.5.

Before proceeding, it is useful to consider the relationship between Bayesian MAP estimates and Tikhonov regularization. From (45) and (46) we can see that in the Gaussian case (or linear MMSE case) the MAP estimate is essentially the same as a general Tikhonov estimate for a particular choice of weighting matrices. For example, suppose that both the noise model and the prior model correspond to uncorrelated random variables so that $\Lambda_q = \lambda_q I$ and $\Lambda_f = \lambda_f I$. Then (45) is equivalent to:

$$\hat{f}_{\text{MAP}} = \arg\min_f \|g - Hf\|_2^2 + \frac{\lambda_q}{\lambda_f} \|f\|_2^2 \tag{50}$$

which is precisely the Tikhonov estimate when $L = I$ and the regularization parameter $\alpha^2 = \lambda_q / \lambda_f$. This association provides a natural interpretation to the regularization parameter as a measure of the relative uncertainty between the data and the prior. As another example, note that the MAP estimate corresponding to the prior model (47) coupled with the observation model (44) with $\Lambda_q = \lambda_q I$ will be the same as a standard Tikhonov estimate with $L = D$ and $\alpha^2 = \lambda_q / \lambda_r$.

While so far the MAP estimate has been interpreted in the Tikhonov context, it is also possible to interpret particular cost choices in the Tikhonov formulation as statistical models for the underlying field and noise. For example, consider

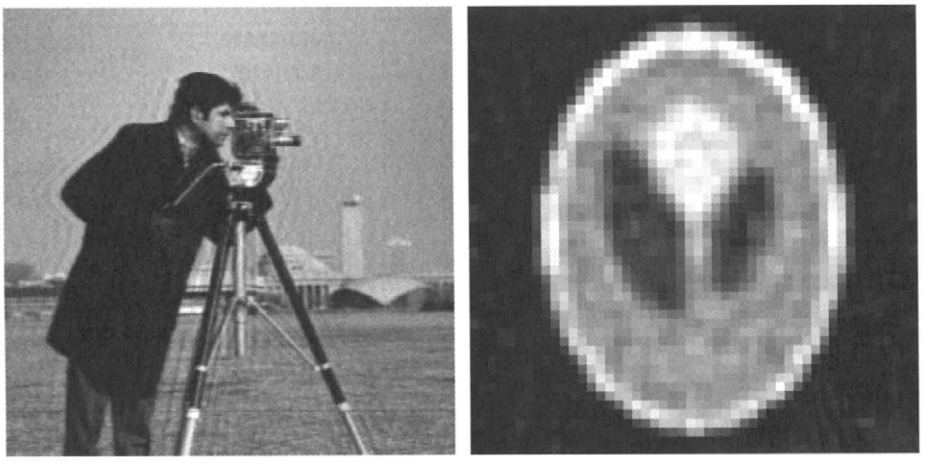


FIGURE 9 Brownian motion prior-based MAP solutions corresponding to the data in Figs. 1(b) (left) and 2(b) (right).

the total variation formulation in (34). Comparing this cost function to (42), it is reasonable to make the following probabilistic associations:

$$\ln p(g \mid f) \propto -\|g - Hf\|_2^2, \qquad \ln p(f) \propto -\alpha^2 \sum_{i=1}^{N_f} \left| [Df]_i \right| \quad (51)$$

which is consistent with the following statistical observation and prior models for the situation:

$$g = Hf + q, \quad q \sim N(0, I) \quad (52)$$

$$p(f) \sim \prod_{i=1}^{N_f} e^{-\alpha^2 \left| [Df]_i \right|}. \quad (53)$$

The statistical prior model for f has *increments* $[Df]_i$ that are independent identically distributed (IID) according to a Laplacian density. In contrast, standard Tikhonov regularization with $L = D$ corresponds to the Brownian motion-type prior model (47), with increments that are also IID but Gaussian distributed.

Finally, while we have emphasized the similarity of the MAP estimate to Tikhonov methods, there are, of course, situations particularly well matched to a statistical perspective. A very important example arises in applications such as in low light level imaging on CCD arrays, compensating for film grain noise, and certain types of tomographic imaging (e.g., PET and SPECT). In these applications, the discrete, counting nature of the imaging device is important and the likelihood term $p(g \mid f)$ is well modeled by a Poisson density function, leading to a signal dependent noise model (see Chapter 4.5). The estimate resulting from use of such a model will be different from a standard Tikhonov solution, and in some instances can be significantly better. More generally, if a statistical description of the observation process and prior knowledge is available through physical modeling or first principle arguments, then the MAP formulation provides a rational way to combine this information together with measures of uncertainty to generate an estimate.

2.5 Parametric Methods

Another direct method for focusing the information in data and turning an ill-conditioned or poorly posed problem into a well-conditioned one is based on changing the parameterization or representation of the problem. The most common representational choice for the unknown image, as mentioned previously, is to parameterize the problem in terms of the values of a regular grid of rectangular pixels. The difficulty with such a parameterization is that there are usually a large number of pixel values which must then be estimated from the available data. For example, an image represented on a 512×512 square pixel array has over 250,000 unknowns to estimate!

Perhaps the simplest change in parameterization is a change in basis. An obvious example is provided by the SVD. By parameterizing the solution in terms of a reduced number of singular vectors we saw that regularization of the resulting solution could be obtained (for example, through TSVD). The SVD is based completely on the distorting operator H, and hence contains no information about the underlying image f or the observed data g. By using parameterizations better matched to these other pieces of the problem it is reasonable to expect better results. This insight has resulted in the use of other decompositions and expansions [14] in the solution of image restoration and reconstruction problems, including the wavelet representations discussed in Chapter 4.2. These methods share the aim of using bases with parsimonious representations of the unknown f, which thus serve to focus the information in the data into a few, robustly estimated coefficients.

Generalizing such changes of basis, prior information can be used to construct a representation directly capturing knowledge of the structure or geometry of the objects in the underlying image f. For example, the scene might be represented as being composed of a number of simple geometric shapes, with the parameters of these shapes taken as the unknowns. For example, such approaches have been taken in connection with problems of tomographic reconstruction [15]. The advantage of such representations is that the number of unknowns can be dramatically reduced to, say, tens or hundreds rather than hundreds of thousands, thus offering the possibility of better estimation of these fewer unknowns. The disadvantage is that the resulting optimization problems are generally nonlinear, can be expensive, and require good initializations to avoid converging to local minima of the cost.

Finally, one class of approaches captures a scene through a series of closed curves, which represent the boundaries of objects in the scene. A cost or energy similar to (30) is formed which depends explicitly on these curves. An estimate is generated by minimizing this cost over the set of possible boundary locations and parameters of the intensities in the corresponding enclosed regions. This minimization is accomplished by evolving the curves so as to reduce the cost. Such "curve evolution" approaches have been applied, for example, to tomography [16]. More discussion of curve evolution methods can be found in Chapter 4.18.

3 Iterative Regularization Methods

One reason for an interest in iterative methods in association with regularization is purely computational. These methods provide efficient solutions of the Tikhonov or MAP normal equations (25) or (46). Their attractions in this regard are several. First, reasonable approximate solutions can often be

obtained with few iterations, and thus with far less computation than required for exact solution of these equations. Second, iterative approaches avoid the memory intensive factorizations or explicit inverses required for exact calculation of a solution, which is critical for very large problems. Finally, many iterative schemes are naturally parallelizable, and thus can be easily implemented on parallel hardware for additional speed.

Interestingly, however, when applied to the *unregularized* problem (12) and terminated long *before convergence,* iterative methods provide a smoothing effect to the corresponding solution. As a result, they can be viewed as a regularization method in their own right [17]. Such use is examined in this section. More detail on various iterative algorithms for restoration can be found in Chapter 3.9. The reason the regularization behavior of iterative algorithms occurs is that the low-frequency (i.e., smooth) components of the solution tend to converge faster than the high-frequency (i.e., rough) components. Hence, for such iterative schemes, the number of iterations plays the role of the inverse of the regularization parameter α, so fewer iterations corresponds to greater regularization (and larger α).

To gain insight into the regularizing behavior of iterative methods, consider a simple Landweber fixed-point iteration for the solution of (12). This basic iterative scheme appears under a variety of names in different disciplines (e.g., the Van Cittert iteration in image reconstruction or the Gerchberg-Papoulis algorithm for bandwidth extrapolation). The iteration is given by:

$$\hat{f}_{\mathrm{lw}}^{(k+1)} = \hat{f}_{\mathrm{lw}}^{(k)} + \gamma H^T\left(g - H\hat{f}_{\mathrm{lw}}^{(k)}\right) = \gamma H^T g + \left(I - \gamma H^T H\right)\hat{f}_{\mathrm{lw}}^{(k)} \tag{54}$$

where γ is a real-valued relaxation parameter satisfying $0 < \gamma < 2/\sigma_{\max}^2$ and $\sigma_{\max}$ is the maximum singular value of H. If the iteration is started with $\hat{f}_{\mathrm{lw}}^{(0)} = 0$, then the estimate after k steps is given by [5, 17]:

$$\hat{f}_{\mathrm{lw}}^{(k)} = \sum_{i=1}^{p}\left[1 - \left(1 - \gamma\sigma_i^2\right)^k\right]\frac{u_i^T g}{\sigma_i}\, v_i \tag{55}$$

where $\{\sigma_i, u_i, v_i\}$ is the singular system of H. Comparing this expression with (19) or (26), the effect of the iterative scheme is again to weight or filter the coefficients of the unregularized generalized solution (19), where the weight or filter function is now given by:

$$w_{i,k} = 1 - \left(1 - \gamma\sigma_i^2\right)^k. \tag{56}$$

This function is plotted in Fig. 10 for $\gamma = 1$ for a variety of values of iteration count k. As can be seen, it has a step-like behavior as a function of the size of σ_i, where the location

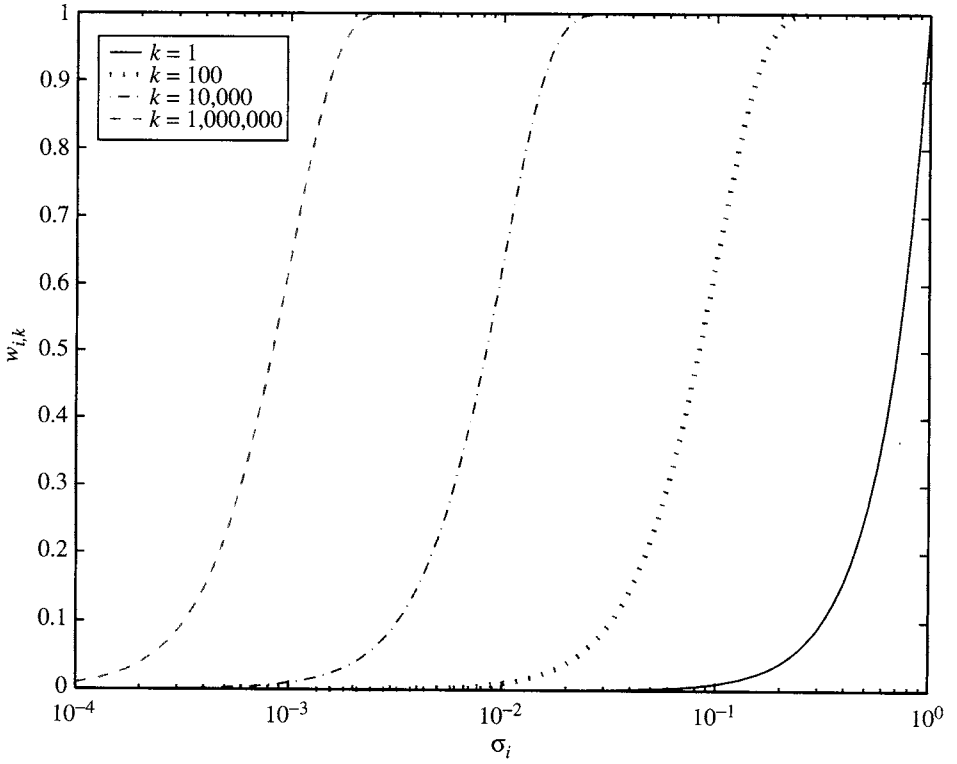


FIGURE 10 Plots of the Landweber weight function $w_{i,k}$ of (56) versus singular value σ_i for various numbers of iterations k.

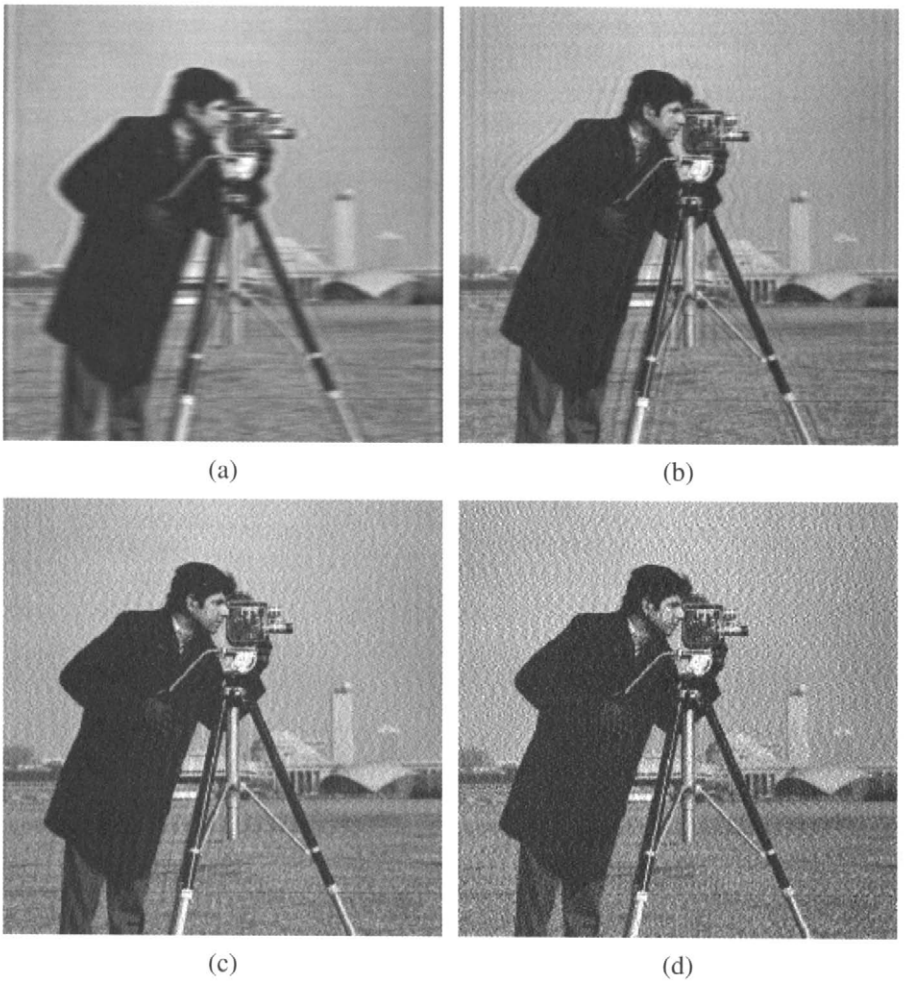


FIGURE 11 Iterative Landweber solution of the unregularized normal equations for the example of Fig. 1. (a) 5 iterations, (b) 50 iterations, (c) 500 iterations, (d) 5,000 iterations.

of the transition depends on the number of iterations [7], so the iteration count of the iterative method does indeed play the role of the (inverse of the) regularization parameter.

An implication of this behavior is that, to obtain reasonable estimates from such an iterative method applied to the unregularized normal equations, a stopping rule is needed, or the generalized inverse will ultimately be obtained. This phenomenon is known as "semi-convergence" [5]. Figure 11 shows Landweber iterative solutions corresponding to the motion-blur restoration example of Fig. 1 after various numbers of iterations and Fig. 12 shows the corresponding solutions for the tomographic reconstruction problem of Fig. 2. In both examples, when too few iterations are performed the resulting solution is over-regularized, blurred, and missing significant image structure. Conversely, when too many iterations are performed, the corresponding solutions are under-regularized and begin to display the excessive noise amplification characteristic of the generalized solution.

While the Landweber iteration (54) is simple to understand and analyze, its convergence rate is slow, which motivates the use of other iterative methods in many problems. One example is the conjugate gradient (CG) method, which is one of the most powerful and widely used methods for the solution of symmetric, sparse linear systems of equations [18]. It has been shown that when CG is applied to the unregularized normal equations (12), the corresponding estimate $\hat{f}_{\mathrm{cg}}^{(k)}$ after k iterations is given by the solution to the following problem:

$$\hat{f}_{\mathrm{cg}}^{(k)} = \arg\min \|g - Hf\|_2 \quad \text{subject to} \quad f \in \mathcal{K}_k\left(H^T H, H^T g\right) \tag{57}$$

where $\mathcal{K}_k(H^T H, H^T g) = \mathrm{span}\{H^T g, (H^T H)H^T g, \ldots, (H^T H)^{k-1} H^T g\}$ is called the Krylov subspace associated to the normal equations. Thus, k iterations of CG again regularizes the problem, this time through the use of a Krylov subspace constraint on the solution (instead of a quadratic side constraint as in (28)). The regularizing effect arises from the property that the Krylov subspace $\mathcal{K}_k(H^T H, H^T g)$ approximates the subspace $\mathrm{span}\{v_1, \ldots, v_k\}$ spanned by the first k right singular

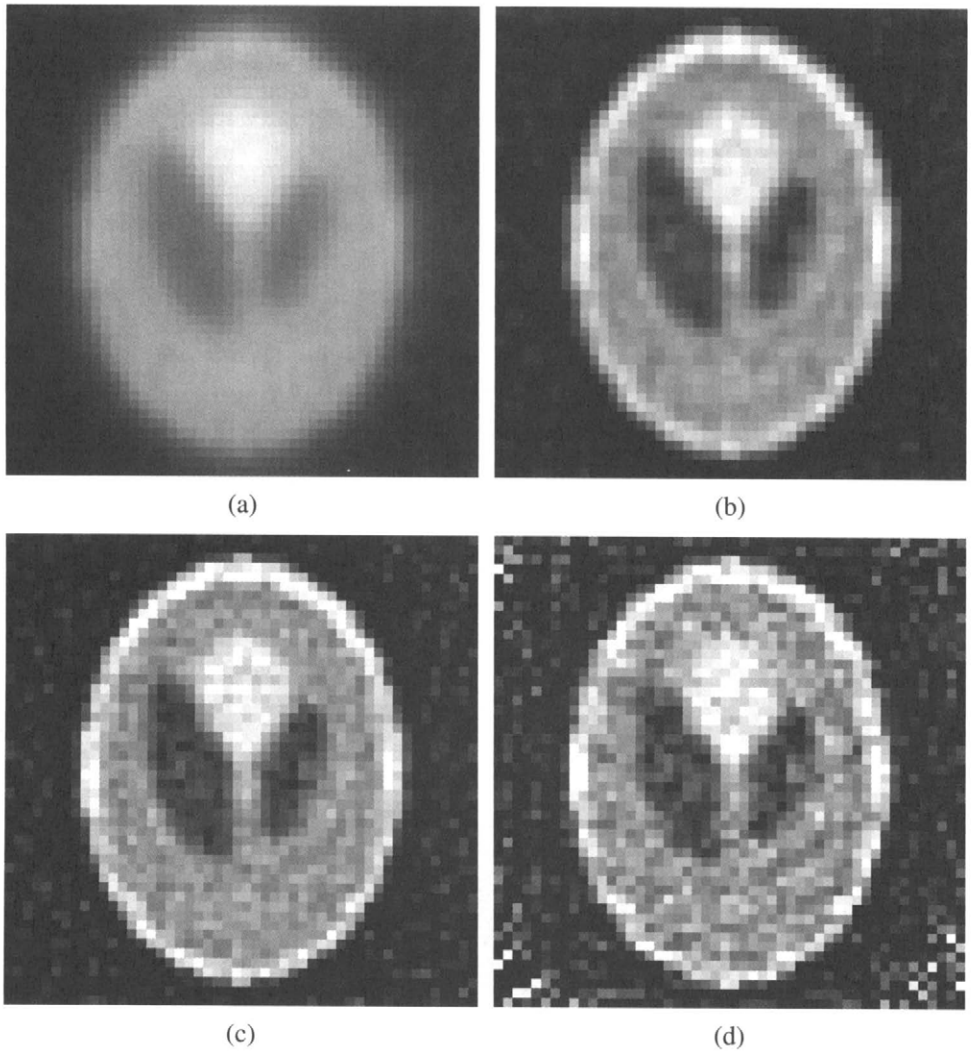


FIGURE 12 Iterative Landweber solution of the unregularized normal equations for the example of Fig. 2. (a) 5 iterations, (b) 50 iterations, (c) 500 iterations, (d) 50,000 iterations.

vectors. While the situation is more complicated than in the Landweber iteration case, weight or filter factors $w_{i,k}$ which depend on k can also be defined for the CG method. These weight factors are observed to have similar attenuating behavior for the small singular values as for the Landweber case, with a roll-off that is also dependent on the number of iterations.

4 Regularization Parameter Choice

Regularization, by stabilizing the estimate in the face of noise amplification, inherently involves a tradeoff between fidelity to the data and fidelity to some set of prior information. These two components are generally measured through the residual norm $\|g - H\hat{f}\|$ (or more generally $J_1(\hat{f}, g)$) and the side constraint norm $\|L\hat{f}\|$ (or more generally $J_2(\hat{f})$). The regularization parameter α controls this tradeoff, and an important part of the solution of any problem is finding a reasonable value for α.

In this section, five methods for choosing the regularization parameter will be discussed: choice based on visual criterion; the discrepancy principle, based on some knowledge of the noise; the L-curve criterion based on a plot of the residual norm versus the side constraint norm; generalized cross-validation, based on minimizing prediction errors; and statistical parameter choice, based on modeling the underlying processes.

4.1 Visual Inspection

Often the main tradeoff dealt with in regularization is between the excessive noise amplification that occurs in the absence of regularization and over smoothing of the solution if too much regularization is used. Further, there may be considerable prior knowledge on the part of the viewer about the

characteristics of the underlying scene — as arises in the restoration of images of natural scenes. In such cases, it may be entirely reasonable to choose the regularization parameter through simple visual inspection of regularized images as the regularization parameter is varied. This approach is well suited, for example, to iterative methods, where the number of iterations effectively sets the regularization parameter. Since iterative methods are terminated long before convergence is achieved when they are used as a form of regularization, the intermediate estimates are simply monitored as the iteration proceeds and the iteration is stopped when noise distortions are observed to be entering the solution. This process can be seen in the examples of Figs. 11 and 12, when few iterations have been performed the solution appears over-regularized. As more iterations are done the detail in the solution is recovered. Finally, as too many iterations are performed the solution becomes corrupted by noise effects. This visual approach to choosing α is clearly problematic in cases where the viewer has little prior understanding of the structure of the scene being imaged or in cases where the reconstructed field itself is very smooth, making it difficult to visually evaluate over- from under-regularized solutions.

4.2 The Discrepancy Principle

If there is knowledge about the perturbation or noise q in (10), then it makes sense to use it in choosing α. When viewed deterministically, this information is often in the form of knowledge about the size or energy of the perturbation:

$$\|q\|_2 \leq \delta_q \tag{58}$$

This knowledge provides a bound on the residual norm $\|g - Hf\|_2 \leq \delta_q$. In a stochastic setting, such information can take the form of knowledge of the noise variance λ_q.

Since the price for over-fitting the solution to the data (i.e., for under-regularizing) is excessive noise amplification (as seen in the generalized solution), it makes sense to choose the regularization parameter large enough that the data fit error achieves this bound, but no larger (to avoid over-regularizing). This idea is behind the discrepancy principal approach to choosing the regularization parameter. Formally, the regularization parameter α is chosen as that value for which the residual norm achieves the equality:

$$\|g - H\hat{f}(\alpha)\|^2 = \delta_q^2 \tag{59}$$

or, in the stochastic setting, where the residual norm equals λ_q. There also exist generalized versions of the discrepancy principle which incorporate knowledge of perturbations to the model H as well. Finally, note that in the deterministic case the value of α provided by the discrepancy principle generally leads to some over-regularization, since the actual perturbation may be smaller than the given bound. Conversely, specification of a bound in (58) that is too small can lead to undesirable noise growth in the solution.

Use of the discrepancy principle requires knowledge of the perturbation bound δ_q or noise variance λ_q. Sometimes this quantity may be obtained from physical considerations, prior knowledge, or direct estimation from the data. When this is not the case, parameter choice methods are required which avoid the need for such knowledge. Two such approaches are examined next.

4.3 The L-Curve

Since all regularization methods involve a tradeoff between fidelity to the data, as measured by the residual norm, and the fidelity to some prior information, as measured by the side constraint norm, it would seem natural to choose a regularization parameter based on the behavior of these two terms as α is varied. Indeed, a graphical plot of $\|L\hat{f}(\alpha)\|_2$ versus $\|g - H\hat{f}(\alpha)\|_2$ on a log-log scale as α is varied is called the L-curve and has been proposed as a means to choose the regularization parameter [7]. Note, especially, that α is a parameter along this curve. The L-curve, shown schematically in Fig. 13 (c.f. [7]) has a characteristic "L" shape (hence its name), which consists of a vertical part and a horizontal part. The vertical part corresponds to under-regularized estimates, where the solution is dominated by the amplified noise. In this region, small changes to α have a large effect on the size or energy of $\hat{f}$, but a relatively small impact on the data fit. The horizontal part of the L-curve corresponds to over-smoothed estimates, where the solution is dominated by residual fit errors. In this region changes to α affect the size of $\hat{f}$ weakly, but produce a large change in the fit error.

The idea behind the L-curve approach for choosing the regularization parameter is that the corner between the horizontal and vertical portions of the curve defines the transition between over- and under-regularization, and thus represents a balance between these two extremes and the best choice of α. The point on the curve corresponding to this α is

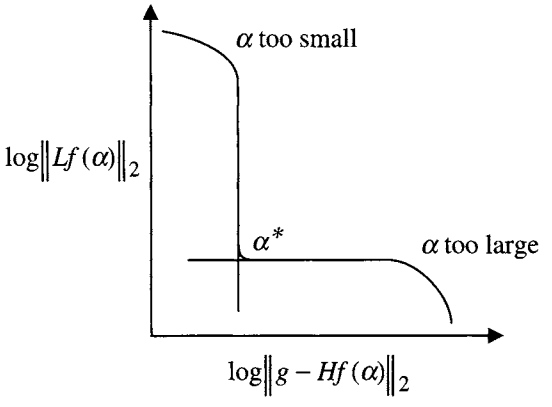


FIGURE 13 Structure of the L-curve.

shown as α^* in Fig. 13. While the notion of choosing α to correspond to the corner of the L-curve is natural and intuitive, there exists the issue of defining exactly what is meant by the "corner" of this curve. A number of definitions have been proposed, including the point of maximum curvature, the point closest to a reference location, and the point of tangency with a line of slope -1. The last definition is especially interesting, since it can be shown that the optimal α for this criterion must satisfy:

$$\alpha^2 = \frac{\|g - H\hat{f}(\alpha)\|_2^2}{\|L\hat{f}(\alpha)\|_2^2}. \tag{60}$$

The right hand side of (60) can be loosely interpreted as the ratio of an estimated noise variance to an estimated signal variance for zero-mean images with $L = I$, and thus appears similar in spirit to (50).

4.4 Generalized Cross-Validation

Another popular method for choosing the regularization parameter that does not require knowledge of the noise properties, is generalized cross-validation (GCV). The basic idea behind cross-validation is to minimize the set of prediction errors — that is, to choose α so that the regularized solution obtained with a data point removed predicts this missing point well when averaged over all ways of removing a point. This viewpoint leads to minimization with respect to α of the following GCV function:

$$\mathcal{V}(\alpha) = \frac{\|g - H\hat{f}(\alpha)\|_2^2}{\left[\text{trace}(I - HH^{\#})\right]^2} \tag{61}$$

where $H^{\#}$ denotes the linear operator which generates the regularized solution when applied to data, so that $\hat{f}(\alpha) = H^{\#}g$. The value of α which minimizes the cost $\mathcal{V}(\alpha)$ in (61) is also an estimate of the value of α which minimizes the mean square error $E[\|Hf - H\hat{f}(\alpha)\|_2^2]$ [19].

Note that only the data is used in the calculation of $\mathcal{V}(\alpha)$ and no prior knowledge of, e.g. the noise amplitude, is required. However, there are also a number of difficulties related to the computation of the GCV cost in (61). First, the operator $HH^{\#}$ must be found. While specifying this quantity is straightforward for e.g. Tikhonov regularization (where it may be written completely in terms of the filter weights $w_{i,\alpha}$ of (27)), it may prove inconvenient for other regularization methods (e.g., for iterative methods; though see [7], Sec. 7.4). Finally, in some cases the GCV cost curve is quite flat, leading to numerical problems in finding the minimum of $\mathcal{V}(\alpha)$, which can result in overly small values of α.

4.5 Statistical Approaches

Our last method of parameter choice is not really a parameter choice technique per say, but rather an estimation approach. As discussed in Section 2.4, given a statistical model of the observation process through $p(g \mid f)$ and of the prior information about f through $p(f)$, the MAP estimate is obtained by solving the optimization problem (42). Note that there are no undetermined parameters to set in this formulation. In the statistical view, the problem of regularization parameter determination is exchanged for a problem of *statistical modeling* through the specification of $p(g \mid f)$ and $p(f)$. The tradeoff between data and prior inherent in the choice of the regularization parameter α is captured in the modeling of the relative uncertainties in the processes g and f. Sometimes the densities $p(g \mid f)$ and $p(f)$ follow from physical considerations or direct experimental investigation. Such is the case in the Poisson observation model for $p(g \mid f)$ often used in tomographic and film-based imaging problems [2, 3]. In such cases, the Bayesian point of view provides a natural and rational way of balancing data and prior.

For many problems, however, the specification of the densities $p(g \mid f)$ and $p(f)$, may appear to be a daunting task. For example, what is the "right" prior density $p(f)$ for the pixels in an image of a natural scene? Identifying such a density at first seems a much more difficult undertaking than finding a good value of the single scalar parameter α (see Chapter 4.7 for a discussion of image densities). Fortunately from an engineering standpoint, the goal is usually not to most accurately model the field f or observation g, but rather to find a *reasonable statistical* model that leads to tractable computation of a good estimate. In this regard, relatively simple statistical models may suffice for the purposes of image restoration and reconstruction. Further, the statistical nature of these models may suggest rational choices of their parameters not obvious from the Tikhonov point of view. For example, as discussed c.f. (50), under a white Gaussian assumption for both the observation noise and the prior, the regularization parameter α^2 can be identified with the variance ratio λ_q/λ_f, where λ_f corresponds to the variance of the underlying image and λ_q is the variance of the noise. Another example is provided by the Brownian motion image model of (47). This case corresponded to Tikhonov regularization with $L = D$ and $\alpha^2 = \lambda_q/\lambda_r$, where now λ_r is the variance of the *derivative* image.

5 Summary

In this chapter we have discussed the need for regularization in problems of image restoration and reconstruction. We have given an overview of the issues that arise in these problems and the means to deal with them. The two driving forces in the need for regularization are noise amplification and lack of data. The primary idea behind regularization is

the inclusion of prior knowledge to counteract these effects. Though there are a large variety of ways to view both these problems and their solution, there is also a great amount of commonality in their essence. In applying such methods to image processing problems (as opposed to 1D signals), all these approaches lead to optimization problems requiring considerable computation. Fortunately, powerful computational resources are becoming available on the desktop of nearly all engineers and there is a wealth of complementary software tools to aid in their application, e.g., [20]. The goal of this chapter has been to provide a unifying view of this area.

Throughout the chapter, we have assumed that the distortion model, as captured by H, is perfectly known and the only uncertainty is due to noise q in the observations g. Often, however, the knowledge of H is not perfect, and in such cases the uncertainty in this model must be dealt with as well. Sometimes it is sufficient to simply treat such model uncertainty as a larger effective observation noise. Alternatively, the uncertainty in H may be explicitly included in the formulation of the inversion problem. Such an approach leads naturally to a method known as total least squares (TLS) [6], which is simply the extension of the least squares idea to include minimization of the square error in both the model and data. Regularized versions of TLS also exist [21], and have shown improved results over the basic least squares methods.

5.1 Further Reading

Chapter 3.5 discusses the basics of image restoration. Chapter 3.7 presents problems arising in multichannel image restoration. Chapter 3.8 treats multi-frame image restoration while Chapter 3.11 is focused on video restoration. In Chapter 3.9 there is a more in depth discussion of iterative methods of image restoration. Chapter 10.2 examines image reconstruction from projections and its application. There are also a number of accessible, yet more extensive, treatments of this material in the general literature. A readable engineering treatment of discrete inverse problems is given in [7] and an associated package of numeric tools is presented in [20]. A deeper theoretical treatment of the topic of data inversion can be found in [4, 5]. The iterative approach to image restoration is studied in [17]. Iterative solution methods in general are discussed in depth in [18].

References

[1] Ranier Kress. *Linear Integral Equations.* Springer-Verlag, New York, 1989.

[2] Avinash C. Kak and Malcolm Slaley. *Principles of Computerized Tomographic Imaging.* IEEE Press, Piscataway, N.J., 1987.

[3] H. C. Andrews and B. R. Hunt. *Digital Image Restoration.* Prentice-Hall, New Jersey, 1977.

[4] M. Bertero. *Advances in Electronics and Electron Physics,* volume 75, chapter Linear Inverse and Ill-Posed Problems, pages 1–120. Academic Press, Boston, 1989.

[5] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems.* Kluwer Academic, Dordrecht, 1996.

[6] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, 1989.

[7] Per Christian Hansen. *Rank-Deficient and discrete Ill-Posed Problems.* Siam, Philadelphia, 1998.

[8] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms." *Physica D,* 60(1–4), 259–268, November 1992.

[9] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. "Maximum entropy and the nearly black object." *J. Roy. Statist. Soc. B,* 54, 41–81, 1992.

[10] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. "Deterministic edge-preserving regularization in computed imaging." *IEEE Trans. Image Proc.,* 6(2), 298–311, February 1997.

[11] Peter J. Huber. *Robust Statistics.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1981.

[12] D. Geman and C. Yang. "Nonlinear image recovery with half-quadratic regularization." *IEEE Trans. Image Proc.,* 4, 932–945, 1995.

[13] A. M. Erisman and W. Tinney. "On computing certain elements of the inverse of a sparse matrix." *Communications of the ACM,* 18(3), 177–179, March 1975.

[14] D. L. Donoho. "Nonlinear solution of linear inverse problems by wavelet-vagulette decomposition." *Applied Computational and Harmonic Analysis,* 2, 101–126, 1995.

[15] P. Milanfar, W. C. Karl, and A. S. Willsky. "Reconstructing binary polygonal objects from projections — A statistical view." *CVGIP: Graphical Models and Image Processing,* 56(5), 371–391, September 1994.

[16] H. Feng, W. C. Karl, and D. A. Castanon. "A curve evolution approach to object-based tomographic reconstruction." *IEEE Trans. Image Processing,* 12, 44–57, January 2003.

[17] Reginald L. Lagendijk and Jan Biemond. *Iterative Identification and Restoration of Images.* Kluwer, Boston, 1991.

[18] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, Cambridge, UK, 1994.

[19] Grace Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[20] Per Christian Hansen. "Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems." *Numerical Algorithms,* 6(1–2), 1–35, 1994.

[21] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary. "Regularization by truncated total least squares." *SIAM J. Sci. Comp.,* 18(4), 1223–1241, July 1997.

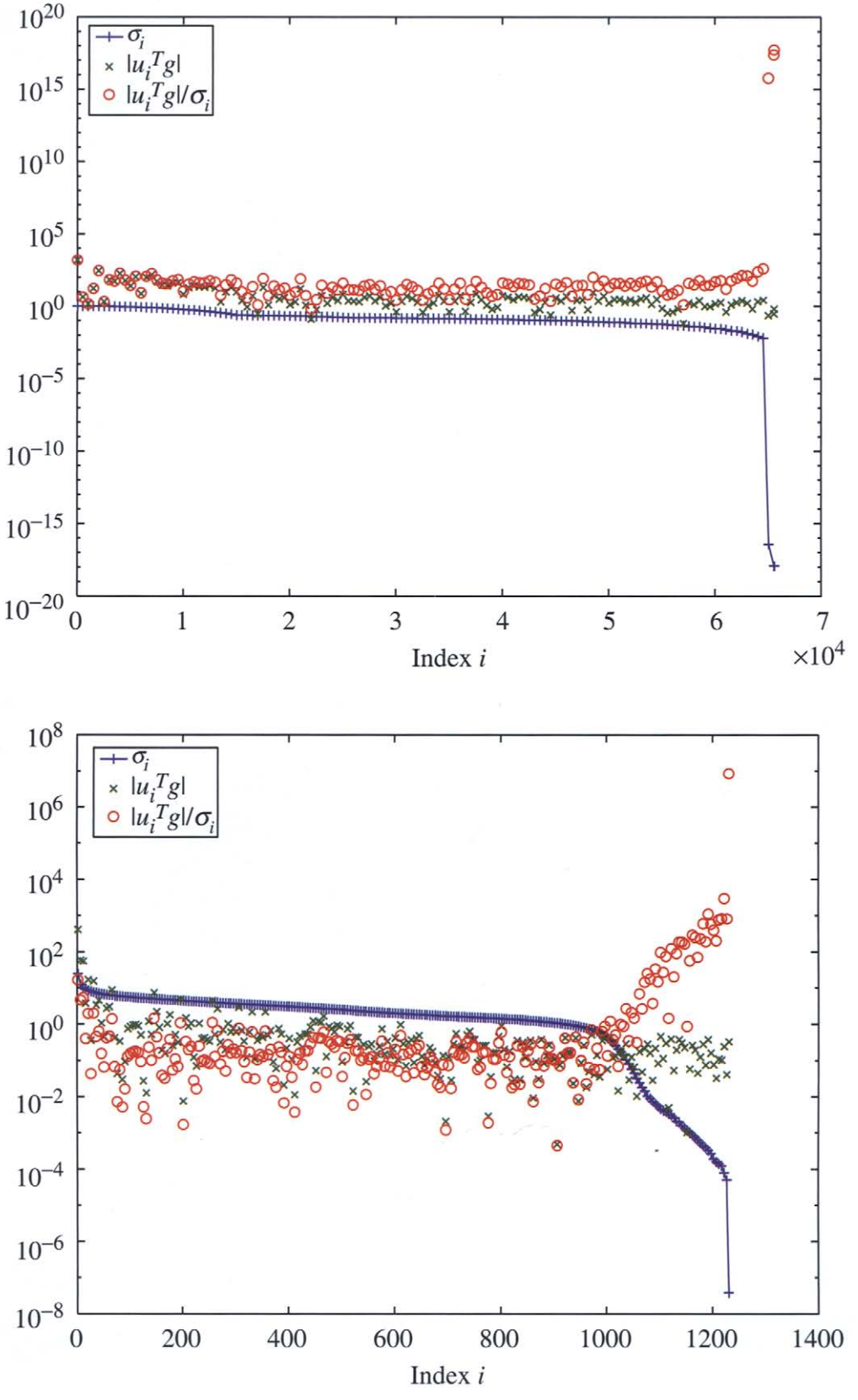


FIGURE 3.6.4 Plots of the components comprising the generalized solution for the problems in Figs. A (top) and B (bottom).