Computing regularization parameters for general-form Tikhonov regularization can be an expensive and difficult task, especially if multiple parameters or many solutions need to be computed in real time. In this work, we assume training data is available and describe an efficient learning approach for computing regularization parameters that can be used for a large set of problems. We consider an empirical Bayes risk minimization framework for finding regularization parameters that minimize average errors for the training data. We first extend methods from Chung et al. (2011 SIAM J. Sci. Comput. 33 3132–52) to the general-form Tikhonov problem. Then we develop a learning approach for multi-parameter Tikhonov problems, for the case where all involved matrices are simultaneously diagonalizable. For problems where this is not the case, we describe an approach to compute near-optimal regularization parameters by using operator approximations for the original problem. Finally, we propose a new class of regularizing filters, where solutions correspond to multi-parameter Tikhonov solutions, that requires less data than previously proposed optimal error filters, avoids the generalized SVD, and allows flexibility and novelty in the choice of regularization matrices. Numerical results for 1D and 2D examples using different norms on the errors show the effectiveness of our methods.

Keywords: spectral filtering, regularization, optimal filters, Tikhonov, learning
1. Introduction

We consider a linear inverse problem that can be modeled as

$$Ax_{\text{true}} + n = b,$$  \hspace{1cm} (1.1)

where $x_{\text{true}} \in \mathbb{R}^{m \times 1}$ is the desired solution, $A \in \mathbb{R}^{m \times n}$, $m \geq n$, models the forward process, $n \in \mathbb{R}^{m \times 1}$ is additive random noise, and $b \in \mathbb{R}^{m \times 1}$ is observed data. The goal is to obtain an approximate solution of $x_{\text{true}}$ given $A$ and $b$. Prior information regarding the probability distribution of $n$ may be incorporated, if known. The matrix $A$ is ill-conditioned and its singular values decay to zero without significant spectral gap. Due to the ill-posed nature of the problem, the exact (inverse) solution of (1.1) will be contaminated by noise and may not be a good approximation of $x_{\text{true}}$. Thus, we seek an approximation to the solution by solving a nearby problem that is well-posed, a process called regularization. A well-known regularization method is the Tikhonov regularization [17]

$$x_\lambda = \arg \min_x \{ \|Ax - b\|_2^2 + \lambda^2 Lx\|_2^2 \},$$  \hspace{1cm} (1.2)

where $\lambda$ is a regularization parameter and $L \in \mathbb{R}^{p \times n}$ is a regularization matrix. We assume that the null spaces of $A$ and $L$ intersect trivially such that $[A^\top L]$ has full row rank and the solution $x_\lambda$ is unique. If $\lambda$ and $L$ are chosen appropriately, the solution to (1.2) should approximate the desired solution $x_{\text{true}}$. Typical choices of $L$ include the identity matrix and the discrete first or second derivative operators. If $L = I$, we say that (1.2) is in standard form. Otherwise, we say it is in general form. For the latter case, the problem can be transformed to standard form. That is, we can use the substitution $x = Ly$ if $L$ is invertible, or otherwise, we can define $x = L_A^\top y$, where $L_A^\top$ is the $A$-weighted generalized inverse of $L$ defined by $L_A^\top = (I - (A(I - L/L)^\top A)^{-1}A)\top L^\top$; see [17, section 2.3] for further details.

The Tikhonov solution $x_\lambda$ can be written as a spectral filtered solution, where both the filter factors and the basis for the solution are determined by the generalized singular value decomposition (GSVD) [18] of $[A^\top L]$. It has been shown that with an appropriate choice of $L$, general-form Tikhonov can produce superior results compared to standard-form Tikhonov [18]. However, for large-scale problems, the difficulty in obtaining the GSVD can present a significant bottleneck towards computing general-form Tikhonov solutions and estimating regularization parameters [1]. Here, we assume that the GSVD is obtainable (e.g., via exploiting structure) or can be efficiently approximated.

Another challenge in previously considered approaches such as iterative techniques for large-scale problems [20, 24, 34] is that an appropriate regularization matrix $L$ must be chosen a priori. We consider multi-parameter Tikhonov, which is an extension that contains several regularization parameters $\lambda = [\lambda_1, \ldots, \lambda_l]$ and multiple regularization matrices $L_j \in \mathbb{R}^{p \times n}$,

$$x_\lambda = \arg \min_x \left\{ \|Ax - b\|_2^2 + \sum_{j=1}^l \lambda_j^2 \|L_j x\|_2^2 \right\}.$$  \hspace{1cm} (1.3)

Previous studies on multi-parameter Tikhonov include [2-4, 13, 28, 36-38].

In this paper, we describe an efficient learning approach for computing regularization parameter(s) that can be used to reconstruct a large set of both general-form Tikhonov (1.2) and multi-parameter Tikhonov (1.3) problems. We assume that training data, whether experimentally obtained or generated via Monte Carlo simulations, is readily available, and we seek regularization parameters that minimize the sample average of errors for the training set. To establish notation and introduce the learning framework, we begin by extending the
work of Chung et al [7] to the general-form Tikhonov problem. Then, a major contribution of this work is to develop an efficient learning approach to compute regularization parameters for the multi-parameter Tikhonov problem, which is more complex, not only because multiple λ’s need to be computed, but also because to the best of our knowledge, there is no extension of the GSVD for multiple L’s where all matrices share the right factor, the left factors are all orthogonal matrices, and the middle factors are diagonal matrices, which is needed to be able to write the solution as a spectral filtered one. Lastly, motivated by our work on the multi-parameter Tikhonov problem, we propose a new class of regularizing filters that encompasses some previously proposed methods and opens the door to new and better filter factors. One advantage of these new filters is that a much smaller training set is required to obtain accurate solutions, compared to the optimal error filter [7]. Additionally, we will show that the regularized solution using the new filter factors corresponds to the solution of a particular multi-parameter Tikhonov problem, but with the advantage that it can be obtained without the GSVD and with flexibility in defining the regularization matrices.

Training data is commonly used in scientific applications such as biomedical and geophysical imaging [10, 12, 33], and previous work on learning approaches in the context of regularization for solving inverse problems can be found in [6, 7, 11, 15, 16, 21, 22, 25, 32]. However, none of these works specifically address the general-form Tikhonov and multi-parameter Tikhonov problems. It is also worth mentioning that in [11, 25], multi-parameter learning approaches for denoising problems (A = I) are proposed. In [25] a parameter learning approach for multiple ℓp-norm regularization terms is presented. The learning problem is formulated as a bilevel optimization problem and solved using semismooth Newton methods. Although the same techniques could be extended to our problem, we believe that such an approach applied to the problem where A ≠ I would be more computationally expensive than the one we propose here.

The paper is organized as follows. Background on general-form Tikhonov regularization is provided in section 2, for readers unfamiliar with the corresponding spectral filtered representation of the solution. In section 3, we describe an empirical Bayes risk framework for computing regularization parameters that are optimal for the training set. We consider both the general-form and multi-parameter Tikhonov problems. We also describe a new regularizing filter that parameterizes the Tikhonov filter factors and show equivalence of the corresponding solution to that of a multi-parameter Tikhonov problem. Numerical results are presented in section 4, and conclusions and future work can be found in section 5.

2. Background

A closed form solution to the general-form Tikhonov regularization problem can be obtained using the GSVD of the matrix pair [A, L],

\[ A = PCZ^{-1} \quad \text{and} \quad L = PSZ^{-1}, \]

where \( P \in \mathbb{R}^{m \times m} \) and \( P \in \mathbb{R}^{p \times p} \) are orthogonal matrices, \( Z \in \mathbb{R}^{n \times n} \) is nonsingular; \( C = \begin{bmatrix} C_A \; 0 \end{bmatrix} \in \mathbb{R}^{m \times n} \), where \( C_A \in \mathbb{R}^{n \times n} \) is diagonal and \( S \in \mathbb{R}^{p \times n} \) is diagonal (not necessarily square), satisfying \( C^T C + S^T S = I \). For convenience, we have ordered the entries in \( C \) and \( S \) such that

\[ 1 \geq c_1 \geq \cdots \geq c_{\min\{n,p\}} \geq 0, \quad c_{\min\{n,p\}+1} = \cdots = c_n = 1, \quad \text{and} \]

\[ L = \begin{bmatrix} D_L \; 0 \end{bmatrix} = \begin{bmatrix} I \; 0 \end{bmatrix}, \]

where \( D_L \in \mathbb{R}^{m \times m} \) and \( D_L \in \mathbb{R}^{p \times p} \) are diagonal.
Using the GSVD, the general-form Tikhonov solution (1.2) can be written as

$$\text{if } i = 1, \ldots, \min\{n, p\},$$

$$(2.2)$$

$$= Z\Phi \left[ \begin{array}{c} C_A^{-1} \\ 0 \end{array} \right] P^* b \equiv A'_i b,$$

$$\text{if } i = \min\{n, p\} + 1, \ldots, n,$$

$$\text{(2.3)}$$

where $\Phi \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal entries are filter factors

$$\phi_i = \begin{cases} \frac{c_i^2}{\sum c_i^2} & \text{if } i = 1, \ldots, \min\{n, p\}, \\ 1 & \text{if } i = \min\{n, p\} + 1, \ldots, n. \end{cases}$$

$$\text{(2.4)}$$

Notice that for $i = \min\{n, p\}$, $\phi_i$ can be written as $\frac{t_i^2}{\sum t_i^2}$, where $t_i = \frac{c_i}{\lambda}$ are the generalized singular values. It is easy to see that for standard-form Tikhonov regularization where $Z$ and $C$ contain the right singular vectors and singular values of $A$ and $s_i = 1$ for all $i$, the filter factors are given by $\phi_i = \frac{c_i^2}{\sum c_i^2 + \lambda^2}$. Therefore, by allowing $L = I$, we not only affect the definition of the filter factors, but also alter the basis vectors $z_i$ used to represent the regularized solution $x_i$.

We assume that the problem satisfies the discrete Picard condition [17]. That is, the values $|p_i|b$ decay on average faster than the values $c_i$, until an index is reached where the noise components dominate the solution. After that point, the coefficients $|p_i|b$ stabilize around the noise level and $c_i$'s continue decreasing, resulting in amplification of errors in the reconstruction. Since the signal is contained primarily in the subspace spanned by $z_i$ for small $i$ and including $z_i$ for larger $i$ results in errors, a good value of $\lambda$ should provide filter factors that filter out the terms in (2.2) corresponding to larger $i$. A sample Picard plot and plot of filter factors are provided in figure 1.

Finding a good regularization parameter is very important, and standard techniques include the discrepancy principle, the L-curve, and the generalized cross-validation (GCV) method [17]. Some of these techniques have been extended to the multi-parameter Tikhonov
problem. For example, a higher dimension L-curve is considered in [3], the discrepancy principle is discussed in [13, 28, 36], an extension of GCV is described in [4], and the Reginiska’s parameter choice rule is generalized in [2]. Many of these methods rely on efficient root finding or optimization methods, which can become extremely costly, especially if multiple parameters need to be computed and for a large set of problems. In the next section, we describe an efficient learning approach for computing regularization parameter(s) and introduce a new class of regularizing filters.

3. Learning regularization parameters from training data

In this section, we describe learning approaches to compute regularization parameters for general-form (1.2) and multi-parameter (1.3) Tikhonov problems for various error measures. We adopt the assumption, common in the machine learning literature, that the training data are samples drawn from a space of solutions according to some probability distribution. We use an empirical Bayes risk minimization framework [5, 22, 26] to incorporate probabilistic information from training data and seek regularization parameters that minimize the sample average of errors.

Suppose we are given training data \( \{(b^k, x^k_{\text{true}}), k = 1, 2, \ldots, K\} \), where \( b^k = Ax_{\text{true}}^k + n^k \), and \( n^k \) is unknown. For each observation \( b^k \), we seek either a general-form Tikhonov solution (1.2) or a multi-parameter Tikhonov solution (1.3) and compute the error vector

\[
e^{k}(\lambda) = x_{\lambda}^{k} - x_{\text{true}}^{k},
\]

where \( \lambda \) is a vector in the multi-parameter case and a scalar in the general-form Tikhonov case. Let \( \rho(\xi) : \mathbb{R}^n \rightarrow \mathbb{R}_+ \) be some error measure, for example, \( \rho(\xi) = ||\xi||_p^p \) for \( p \geq 1 \). Then the goal is to find optimal regularization parameters \( \lambda \) that minimize the average error of the reconstructions for the training data

\[
f_K(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \rho(e^{k}(\lambda)).
\]

That is, we seek a solution to the following empirical Bayes risk minimization problem

\[
\min_{\lambda} f_K(\lambda).
\]

Notice that underlying (3.3) is a Bayes risk minimization problem, where the goal is to compute parameters \( \lambda \) that minimize the expected value of the reconstruction errors over the joint distribution of the noise and unknown parameters [7]. Since these distributions are rarely known in practice, we approximate the expected value with the sample mean and instead seek a solution to (3.3). Minimizing the average error (3.2) is one approach based on this interpretation that we are minimizing the expected value of the errors, but one could also consider minimizing the median of the errors or the maximum value of the errors, depending on the desired design or goal of the problem [7, 15].

3.1. One-parameter general-form Tikhonov

For the one-parameter general-form Tikhonov problem, derivative-free methods (e.g., \texttt{fminsearch} in MATLAB) could be used to solve (3.3) for various error measures. However, since the derivative \( f'_K(\lambda) \) can be easily computed when the GSVD is available,
standard root finding methods can also be used to solve $f'_k(\lambda) = 0$, see e.g., [29]. For completeness, we provide the derivatives here.

Let us define vector $\gamma \in \mathbb{R}^{n \times 1}$ with elements $\gamma_i = \frac{u_i}{e_i}$ and let $\Gamma \in \mathbb{R}^{n \times n}$ be a diagonal matrix with entries $\gamma_i$. Then partition

$$
Z = \begin{bmatrix} \tilde{Z}, & \hat{Z} \end{bmatrix}, \quad \gamma = \begin{bmatrix} \gamma \\hat{\gamma} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \tilde{\Gamma} & & & \\ & \hat{\Gamma} \end{bmatrix},
$$

(3.4)

where $\tilde{Z} \in \mathbb{R}^{m \times \min\{n,p\}}$, $\hat{\gamma} \in \mathbb{R}^{\min\{n,p\} \times 1}$, $\hat{\Gamma} \in \mathbb{R}^{\min\{n,p\} \times \min\{n,p\}}$, and $\tilde{Z}, \hat{\gamma}, \hat{\Gamma}$ have corresponding size. Then, the general-form Tikhonov solution (2.3) can be written as

$$
x^{(k)} = \sum_{i=1}^{n} \phi_i \frac{p_i e^{(k)}}{c_i} z_i = \tilde{Z} \hat{\Gamma} \phi(\lambda) + \tilde{Z} \hat{\gamma},
$$

(3.5)

where $\phi \in \mathbb{R}^{\min\{n,p\} \times 1}$ contains filter factors $\phi_i$ which depend on $\lambda$. For each $k$, the Jacobian of the errors with respect to $\lambda$ can be written as

$$
J^{(k)} = \tilde{Z} \hat{\Gamma}^{(k)} \psi,
$$

(3.6)

where the $i$th element of vector $\psi$ is the derivative of $\phi_i$ with respect to $\lambda$, given by

$$
\psi_i = -\frac{2 \lambda e_i^2 + \lambda}{(\lambda e_i^2 + \lambda)^2}.
$$

Thus, the derivative $f'_k(\lambda)$ can be computed as

$$
f'_k(\lambda) = \frac{1}{K} \sum_{k=1}^{K} J^{(k)} \nabla \rho(e^{(k)}),
$$

(3.7)

where $\nabla \rho$ contains the partial derivatives of $\rho$ evaluated at $e^{(k)}(\lambda)$. We remark that the second derivative, if desired, could be numerically approximated.

Henceforth, we use ‘opt-Tik-GSVD’ to denote general-form Tikhonov solutions where the regularization parameter is optimal for the training set and ‘opt-Tik-SVD’ to denote standard-form Tikhonov solutions where the regularization parameter is optimal for the training set.

For one realization, we consider how far the general-form Tikhonov solution that uses a learned parameter can be from the solution using an optimal parameter. The following result gives an upper bound.

**Theorem 3.1.** Let $\lambda_{\text{opt}}$ be the optimal parameter for the Tikhonov problem, that is, optimal in the sense that it corresponds to the Tikhonov solution (1.2) closest to $x_{\text{true}}$. If $L$ is square and invertible, then for any value of $\lambda$,

$$
\frac{\|x_{\text{true}} - x\|_2}{\|x_{\text{true}}\|_2} \leq \text{cond}(L) \frac{|\lambda_{\text{opt}} - \lambda|}{t^* + \lambda^2},
$$

(3.8)

where $t_*$ is the smallest generalized singular values of the pair $\{A, L\}$.

**Proof.** The proof follows from applying theorem 3.1 in [31] to the equivalent standard-form Tikhonov problem and the fact that the singular values of $AL^{-1}$ are the generalized singular values of the pair $\{A, L\}$. \hfill $\square$

In [7], optimal error filters were introduced, where no functional form for the filter factors was assumed. Analogous filters can be defined for general-form Tikhonov. That is, the reconstruction has the form (3.5), where there are $n$ regularization parameters, $\lambda_i$, and filter factors $\phi_i = \lambda_i$ are optimal in that they minimize the average reconstruction error for the
training data. We refer to these solutions as ‘opt-error-GSVD’. In standard-form Tikhonov, the reconstruction is written in terms of the right singular vectors of \( \mathbf{A} \), and we refer to these solutions as ‘opt-error-SVD’. One of the caveats of the optimal error filters is that since \( n \) filter factors need to be computed, more training data is required to avoid potential overfitting or bias to the training set. This will be investigated in section 4.

3.2. Multi-parameter Tikhonov

In this section we describe a learning approach to compute regularization parameters for the multi-parameter Tikhonov problem (1.3), whose solution can be written as

\[
x_{\lambda} = \left( \mathbf{A}^\top \mathbf{A} + \sum_{j=1}^{J} \lambda_j^2 \mathbf{L}_j \right)^{-1} \mathbf{A}^\top \mathbf{b}.
\]  

(3.9)

Unlike the one-parameter Tikhonov problem where the GSVD provides a decomposition that allows the solution to have filtered representation (2.2), such a factorization where all matrices share the same right factor, the left factors are orthogonal matrices, and the middle factors are diagonal matrices

\[
\mathbf{A} = \mathbf{P} \mathbf{C} \mathbf{Z}^{-1} \quad \text{and} \quad \mathbf{L}_j = \tilde{\mathbf{P}}_j \mathbf{S} \mathbf{Z}^{-1} \quad j = 1, \ldots, J,
\]

(3.10)

may not exist in general [4]. Thus, a filtered representation for \( x_{\lambda} \) is not always possible. For problems where such a decomposition does not exist, a bilevel optimization approach could be used to compute regularization parameters [9, 25]. However, this approach would require computing (3.9) for various estimates of \( \lambda \), which could become computationally expensive. Another approach that was proposed in [8] uses operator approximations to estimate regularization parameters. In particular, for problems where operator approximations can be found that satisfy the above simultaneous diagonalizability condition (3.10), we investigate a framework where the approximate problem is used to estimate near-optimal regularization parameters and then the original problem is solved using the computed parameters. Numerical investigations for the multi-parameter case are provided in example 2 in section 4.

Next, we restrict our derivation to square matrices, i.e., \( m = n \) and \( p_j = n \) for \( j = 1, \ldots, J \), and consider scenarios where such a decomposition exists or can be approximated.

In spatially invariant signal and image processing, it is common to encounter matrices \( \mathbf{A} \) and \( \mathbf{L}_j \) that satisfy (3.10), under certain assumptions on the boundary conditions. In these situations, \( \mathbf{P}, \tilde{\mathbf{P}}_j \), and \( \mathbf{Z}^{-1} \) typically all represent the same frequency transform, and we get the following simultaneous diagonalizability condition

\[
\mathbf{A} = \mathbf{Q} \mathbf{C} \mathbf{Q}^* \quad \text{and} \quad \mathbf{L}_j = \mathbf{Q} \mathbf{S} \mathbf{Q}^* \quad j = 1, \ldots, J,
\]

(3.11)

where \( \mathbf{Q} \) is an orthogonal (or unitary) matrix. For instance, in image deblurring, if we assume periodic boundary conditions on the image, then the blurring operator \( \mathbf{A} \) and the regularization matrices \( \mathbf{L}_j \) are block circulant matrices with circulant blocks. In this case, \( \mathbf{Q}^* \) represents the 2D discrete Fourier transform (DFT) matrix. If we assume reflexive boundary conditions and the point spread functions defining \( \mathbf{A} \) and \( \mathbf{L}_j \) satisfy a double symmetry condition [19], then the matrices can each be written as a sum of BTTB (block Toeplitz with Toeplitz blocks), BTHB (block Toeplitz with Hankel blocks), BHBB (block Hankel with Toeplitz blocks), and BHHB (block Hankel with Hankel blocks) matrices. In this case, \( \mathbf{Q}^* \) represents the 2D discrete cosine transform (DCT) matrix.

Next, we show that with decomposition (3.11), the multi-parameter Tikhonov solution can be written as a filtered solution. For notational purposes, \( |\Delta|^2 \) represents element-wise absolute value followed by element-wise square for any matrix \( \Delta \). From (3.9), the solution
of (1.3) can be written as
\[
x_{\lambda} = Q \left( |C|^2 \left( |C|^2 + \sum_{j=1}^{J} \lambda_j |S_j|^2 \right)^{-1} \right) C^{-1}Q^*b
\]
\[
= \sum_{i=1}^{n} \phi_i q_i^* b_i
\]
\[
= Q\Phi C^{-1}Q^*b,
\]
where as before $\Phi$ is a diagonal matrix containing the filter factors, which in this case are given by
\[
\phi_i = \frac{|c_i|^2}{|c_i|^2 + \sum_{j=1}^{J} \lambda_j |s_{ij}|^2},
\]
with $s_{ij}$ being the $i$th diagonal element of matrix $S_j$.

For notational convenience, let $\Gamma$ be an $n \times n$ diagonal matrix whose $i$th diagonal element is $\gamma_i = \frac{q_i^* b_i}{s_i}$. Then (3.12) can be written as
\[
x_{\lambda} = Q\Gamma \phi(\lambda),
\]
where $\phi(\lambda) \in \mathbb{R}^{n \times 1}$ is the vector of filter factors (3.13) that depends on the vector of regularization parameters $\lambda$.

Recall that our goal is to compute regularization parameters that are optimal for the training data on average by solving (3.3), where instead of one regularization parameter $\lambda$, we now have multiple parameters $\lambda_1, \ldots, \lambda_J$, and therefore the error in (3.2) is given by
\[
\epsilon^{(k)}(\lambda) = Q\Gamma^{(k)} \phi(\lambda) - x_{\text{true}}^{(k)}.
\]
Here, for each training data $b^{(k)}$, we define $\Gamma^{(k)}$ where the $i$th diagonal element of $\Gamma^{(k)}$ is $\gamma_i^{(k)} = \frac{q_i^* b_i^{(k)}}{s_i}$.

Since we must optimize over several parameters, we propose to use a Gauss–Newton approach to solve (3.3). The Jacobian of the errors with respect to $\lambda$ can be written as
\[
J = \frac{1}{K} \begin{bmatrix} J^{(1)} \\ \vdots \\ J^{(K)} \end{bmatrix},
\]
with $J^{(k)} = Q\Gamma^{(k)}\Psi$, (3.16)
where the $(i,j)$th entry of the matrix $\Psi \in \mathbb{R}^{n \times J}$ is the derivative of $\phi_i$ with respect to $\lambda_j$. Define $S \in \mathbb{R}^{n \times J}$ whose $i,j$th entry is $|s_{ij}|^2$, then the matrix $\Psi$ is given by
\[
\Psi = -2TS\Lambda,
\]
where $\Lambda = \text{diag} \lambda$ and $T$ is a diagonal matrix whose $i$th diagonal entry is given by $|c_i|^2 + \sum_{j=1}^{J} \lambda_j |s_{ij}|^2$. The gradient of $f_k(\lambda)$ and the Gauss–Newton approximation of the Hessian can be written as
\[
g = \frac{1}{K} \sum_{k=1}^{K} J^{(k)} \nabla_{\epsilon^{(k)}} \rho(\epsilon^{(k)}), \quad \text{and} \quad H = \frac{1}{K} \sum_{k=1}^{K} J^{(k)} \nabla_{\epsilon^{(k)}}^2 \rho(\epsilon^{(k)}),
\]
where $\nabla_{\epsilon^{(k)}} \rho(\epsilon^{(k)})$ contains the partial derivatives of $\rho$ evaluated at $\epsilon^{(k)}(\lambda)$, and $\nabla_{\epsilon^{(k)}}^2 \rho(\epsilon^{(k)})$ contains the second derivatives (i.e., the Hessian) of $\rho$ evaluated at $\epsilon^{(k)}(\lambda)$. 
For the special case where \( \rho(\xi) = \frac{1}{2} \| \xi \|_2^2 \), we have \( \nabla^2_{\theta_i}(\hat{\epsilon}(\theta_i)) = \hat{\epsilon}(\theta_i) \) and \( F^{(k)} = I \), so the Hessian approximation simplifies nicely

\[
H = \frac{1}{K} \sum_{k=1}^{K} \hat{J}^{(k)} \hat{J}^{(k)} = \frac{1}{K} \sum_{k=1}^{K} \Psi^* (\Gamma^{(k)})^* Q^* Q \Gamma^{(k)} \Psi \\
= \frac{4}{K} \bar{A} \bar{S}^* \left\{ T \left( \sum_{k=1}^{K} |\Gamma^{(k)}|^2 \right) T \right\} \bar{S} A,
\]

where we note that the matrix contained in the curly brackets is diagonal. Notice that for the case \( J = 1 \), these derivations are equivalent to those for the one-parameter general-form Tikhonov case.

### 3.3. A new class of regularizing filters

In this last subsection, we propose a new class of regularizing filters that is motivated by our above work on multi-parameter Tikhonov regularization, but that can be used for general filtering. Let \( \lambda \in \mathbb{R}^{n \times 1} \) be a vector of regularization parameters and define \( \bar{X}^2 \) to be the element-wise squared vector. Then we propose to use Tikhonov-like filter factors of the form

\[
\phi(\lambda) = \frac{\sigma_i^2}{\sigma_i^2 + (S \bar{X})^2},
\]

where \( \sigma_i \) is the \( i \)th singular value of \( A \) and \( (S \bar{X})_i \) is the \( i \)th element of the vector \( S \bar{X}^2 \), where the columns of user-defined matrix \( S \in \mathbb{R}^{n \times J} \) provide a subspace for the regularization parameters. For example, if \( S \) is a column of ones, then the filter factors reduce to the standard-form Tikhonov filter factors. If \( \sigma_i = c_i \) and \( S \) is a column containing entries \( s_j^2 \) from the GSVD \( (2.1) \) then we get the general-form Tikhonov filter factors.

By noticing the connection to the derivation in section 3.2, it is easy to see that filter factors \( (3.19) \) correspond to a multi-parameter Tikhonov problem with \( \bar{L}_j = \bar{U}_j D_j \bar{V} \) where \( \bar{U}_j \) can be any orthogonal matrix, \( \bar{V} \) comes from the SVD of \( A \), and \( D_j \) is a diagonal matrix with the element-wise square root of the \( j \)th column of \( S \) on the diagonal. Thus, we know that the solution using this regularizing filter corresponds to that of a multi-parameter Tikhonov problem. However, we do not require the GSVD and there is flexibility and potential novelty in the definition of the subspace vectors used to define matrices \( \bar{L}_j \). Furthermore, numerical results illustrate that reconstructions are of similar quality to that of the optimal error filter but with the advantage that it requires fewer training data.

Parameterizing filter factors is common practice. For example, TSVD and Tikhonov filters have well-studied functional forms, spline filters were proposed in [7], and other regularizing filters were considered in [35]. However, our general formulation is different in that it allows for any number of regularization parameters and any given subspace. Some examples of subspaces include polynomials, splines, or a subspace consisting of eigenvalues from derivative matrices.

As described above, our proposed filter encompasses some previously proposed filters, but it also opens the door to the development of new filters. Therefore, we refer to this new family of filter factors as subspace-Tikhonov (S-Tik) filter factors.
4. Numerical results

In this section, we provide some numerical results that illustrate the performance of the learned regularization parameters for general-form and multi-parameter Tikhonov regularization. Following convention in [7], we refer to these parameters as ‘optimal’ in the sense that they are optimal for the training data on average. We consider three investigations. Example 1 is a 1D deconvolution example that compares optimal Tikhonov and optimal error filters for both standard SVD and generalized SVD. Examples 2 and 3 are 2D deconvolution examples. Example 2 compares one versus multiple \( L \)'s for various error measures, and investigates using approximate matrices to learn regularization parameters for the original problem. Then example 3 investigates the benefits of the new S-Tik filter, compared to standard filters.

Example 1. Here we compare general-form to standard-form Tikhonov in the optimal filter framework. In this example, we investigate their performance in the deconvolution of 1D signals. To generate training signals, we took 200 columns of each of the 5 MRI images shown at the top of figure 2, resulting in 1000 training signals, each of size \( 256 \times 1 \). Each signal was blurred with a Gaussian point spread function with mean 0 and variance 1. Gaussian white noise was added, where noise levels were randomly selected between 0.2 and 0.25. That is, a noise level of 0.2 means \( \|n\|^2/\|Ax_{true}\|^2 = 0.2 \). Another 1000 signals were generated for validation in the same manner but extracted from 5 different MRI images shown at the bottom of figure 2.

For this example, we select regularization matrix \( L \in \mathbb{R}^{257 \times 256} \) to represent a discretization of a first derivative operator.
We compare four approaches that use training signals to obtain optimal filters.

- **opt-Tik-SVD** and **opt-Tik-GSVD** correspond to standard-form and general-form Tikhonov solutions, where $\lambda$ minimizes the average mean squared error over all training data, i.e., $\lambda$ solves (3.3) with $\rho(\xi) = \|\xi\|^2$.
- **opt-error-SVD** and **opt-error-GSVD** correspond to solutions that use the optimal error filters for the SVD and GSVD bases, respectively. That is, no functional form is assumed for the optimal error filter, so optimization is performed over 256 unknowns.

It is worth noting that opt-Tik-SVD and opt-error-SVD filters were introduced and compared in [7]. Our goal here is to investigate their GSVD counterparts.

Once computed, the optimal filters were used to reconstruct each validation signal. The distribution of the relative reconstruction errors (RREs), $\|\hat{x}_i - x_{\text{true}}\|^2_2/\|x_{\text{true}}\|^2_2$, for the validation set is presented in figure 3 using box-and-whisker plots. The box part presents the median value, along with the 25th and 75th percentiles. The whiskers correspond to extreme data points, and outliers are plotted individually. As observed in [7], relative errors for opt-Tik-SVD are consistently higher than those for opt-error-SVD due to the high noise level in the problem. The optimal Tikhonov filter corresponding to GSVD can produce errors that are comparable to both opt-error-SVD and opt-error-GSVD. This is an important point because both opt-error-SVD and opt-error-GSVD require $n$ regularization parameters, whereas opt-Tik-GSVD only requires one. Of course, the performance of opt-Tik-GSVD relies on a good choice of $L$; such concerns will be addressed in later examples of the multi-parameter case.

Filter factors for the four approaches are presented in figure 4. In the top plot, we observe that opt-Tik-SVD has difficulty in obtaining filter factors close to 1 (corresponding to large singular values). Since Tikhonov filter factors depend on one regularization parameter, there is an inherent limitation of standard-form Tikhonov for problems with high noise levels. However, this is not true of Tikhonov filter factors with GSVD, as evident in the bottom plot in figure 4. Thus, one of the potential benefits of using general-form Tikhonov is being able to obtain filter factors that can overcome the limitations of standard-form Tikhonov.

\[
\mathbf{L} = \begin{pmatrix}
1 & 1 \\
-1 & 1 \\
-1 & 1 \\
-1 & 1
\end{pmatrix}
\]
validating previous observations (e.g., in [18]) of the superiority of general-form over standard-form Tikhonov solutions.

Lastly, for this example, we investigate reconstruction errors for the validation set as a function of the number of training signals used to obtain the optimal filters. We provide average reconstruction errors for the validation set in figure 5. These plots are often referred to as Pareto curves. We see that average RREs for opt-Tik-SVD and opt-Tik-GSVD remain fairly stable as the number of training signals increases. On the other hand, for small numbers of training signals, opt-error-SVD and opt-error-GSVD are significantly biased towards the

Figure 4. Computed filter factors for approaches compared in example 1.

Figure 5. Pareto curves that display the average relative reconstruction errors for the validation set as a function of the number of training signals used to compute the optimal filters in example 1.
training data, but reconstruction errors plateau with enough training signals. It is interesting to note that opt-error-SVD requires 298 training signals to achieve the same RRE that opt-Tik-GSVD can obtain with only 1 training signal. Thus, opt-Tik-GSVD can be a good alternative to opt-error-SVD if a good $L$ is known a priori and if there are very few signals in the training set.

**Example 2.** The goal of this example is to use a 2D image deblurring problem to compare one versus multi-parameter Tikhonov in the learning framework, as well as to compare different error measures $\rho$. We also compare to standard parameter selection methods and consider a problem where approximate matrices are used to learn regularization parameters for the original problem. Since we only consider Tikhonov solutions in this example, ‘Tik’ is removed from labels for clarity of presentation.

For the training set, we used eight satellite images of size $256 \times 256$ with 10 rigid transformations of each, for 80 total training images. A 2D symmetric Gaussian blur with zero mean and variance 1 was used, and reflexive boundary conditions were assumed. Additive white Gaussian noise was included with a noise level that was randomly selected from a range of 0.1 and 0.15. A set of 80 validation images was generated in the same manner, but with eight different satellite images. Sample training and validation images can be found in the top and bottom rows of figure 6, respectively.

We consider various choices for regularization matrix $L$. We use $L_1 \equiv I$ (i.e., standard-form Tikhonov), finite difference matrices that approximate the second derivative in each direction, denoted $L_2$ and $L_3$, and the discrete Laplacian matrix, $L_4$. Stencils corresponding to the latter three matrices are given by

$$
\begin{bmatrix}
0 & 0 & 0 \\
1 & -2 & 1 \\
0 & 0 & 0
\end{bmatrix},
\begin{bmatrix}
0 & 1 & 0 \\
0 & -2 & 0 \\
0 & 1 & 0
\end{bmatrix}, \quad \text{and} \quad
\begin{bmatrix}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 0
\end{bmatrix}
$$

(4.2)

Since all of these stencils are doubly symmetric and we assumed reflexive boundary conditions, simultaneous diagonalizability (3.11) can be obtained where $Q^\alpha$ represents the 2D DCT matrix.
Using $\rho(\xi) = \|\xi\|^2_2$, we used the training images to compute optimal regularization parameters for each $L_i$ individually, which we refer to as opt-L1, opt-L2, opt-L3, and opt-L4 respectively. In addition, optimal parameters were computed for the multi-parameter Tikhonov problem using all four $L_i$ matrices; we call this opt-multi. RREs, computed using the 2-norm, for the validation images are presented in figure 7 using box plots. We note that the Laplacian matrix provided smaller relative errors than the other $L_i$'s, and multi-parameter Tikhonov consistently provided good results.

Similar results were observed for other error functions. In particular, we considered $\rho(\xi) = \|\xi\|^2_1$ and $\rho(\xi) = \|\xi\|_1$. Since optimization of the 1-norm can be difficult due to non-differentiability, we use the Huber function [23]

$$\rho(\xi) = \begin{cases} \frac{\beta}{2} |\xi| - \frac{\beta}{2}, & \text{if } |\xi| \geq \beta, \\ \frac{1}{2\beta} \xi^2, & \text{if } |\xi| < \beta, \end{cases} \quad (4.3)$$

with smoothing parameter $\beta = 10^{-4}$, which provides an approximation to the 1-norm. RREs were computed as $\rho(x_i - x_{\text{true}})/\rho(x_{\text{true}})$. The average RRE for the validation set and corresponding standard deviations can be found in table 1. For comparison, we provide results that correspond to using the well-known GCV method [14] to select parameters for multi-parameter Tikhonov, denoted GCV-multi. That is, for each validation image, GCV parameters $\lambda_{\text{GCV}}$ are computed to minimize the GCV function

$$\text{GCV}(\lambda) = \frac{\| (I - AA_0^*)b \|^2_2}{\text{trace}(I - AA_0^*))^2} = \frac{\sum_{j=1}^n (1 - \phi_j)^2 (q_j^*b)^2}{\left(\sum_{j=1}^n (1 - \phi_j)^2\right)^2}, \quad (4.4)$$

where $A_0^* = \Phi \Phi^* \Phi^{-1} \Phi^*$. Notice that for this example, we can exploit simultaneous diagonalizability to obtain a multidimensional representation of the GCV function that is easy to evaluate. However, in general, this approach to computing regularization parameters for a large set of problems can become prohibitively costly, as it may require many multi-parameter Tikhonov solves within an optimization scheme. For such cases, an alternative proposed in [4] is to approximate the solution of the multi-parameter Tikhonov problem by a weighted sum of one-parameter Tikhonov solutions, where each regularization parameter is selected independently using 1D GCV functions. Although such an approach is not necessary...
for our problem, we provide results using this approach, which we denote as GCV1D-multi, where again we were able to use simultaneous diagonalizability to simplify the evaluation of the 1D GCV functions. For simplicity, we use a simple average of the, here, four regularized solutions but remark that more accurate solutions may be obtained by using the scheme described in [4].

We observe that in all error measures, the Laplacian outperformed all of the tested $L$ matrices, even multi-parameter Tikhonov. Note that for the 2-norm, GCV-multi performs very well, even better than opt-multi, which is expected since opt-multi computes regularization parameters that are optimal on average, while GCV-multi computes regularization parameters that are specific to each realization. However, this is not the case for the other error measures.

The choice of $\rho$ is reflected in the reconstructed images. For one validation image, we provide in figure 8 absolute error images (i.e., absolute value of the reconstruction minus the true image) in inverted colormap. White regions in the image correspond to low absolute errors and darker regions correspond to larger errors. Consistent with the above observations, we see that for all considered error functions, opt-$L_4$ and opt-multi provide reconstructions with smaller absolute errors than opt-$L_3$, opt-$L_2$, and opt-$L_1$. Comparing different error functions, we observe that reconstructions with Huber have small errors in flat or constant regions of the image and large errors near edges. Furthermore, 5-norm reconstructions have overall smaller reconstruction errors, but the errors are distributed in the image. This is consistent with observations made in [7]. The error images on the far right correspond to Tikhonov reconstructions using GCV.

Next we investigate the use of operator approximations to estimate regularization parameters in our learning framework. As mentioned in section 3.2, for problems where simultaneous diagonalizability does not exist or is not easily obtainable, e.g., for spatially variant blurs [30] or problems where $L$ represents dictionaries or gradient masks [27], a bilevel optimization approach could be used, but may be costly. Instead, we use operator approximations [8] to estimate regularization parameters from the training data, and then solve the original problem using these parameters.

In order to validate our results, we use the 2D image deblurring example, where all involved matrices $A$, $L_1$, $L_2$, $L_3$, and $L_4$, are simultaneously diagonalizable by the 2D DCT. However, we assume that such a factorization is not available, and we use the assumption of periodic boundary conditions to obtain approximate matrices $\tilde{A}$, $\tilde{L}_1$, $\tilde{L}_2$, $\tilde{L}_3$, and $\tilde{L}_4$, which are

### Table 1. Average relative reconstruction error and standard deviation, in parenthesis, for the validation set in example 2 for various error functions.

<table>
<thead>
<tr>
<th></th>
<th>Huber</th>
<th>2-norm</th>
<th>5-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>opt-$L_1$</td>
<td>1.672e-01</td>
<td>2.324e-02</td>
<td>1.232e-04</td>
</tr>
<tr>
<td>opt-$L_2$</td>
<td>3.718e-01</td>
<td>1.257e-01</td>
<td>1.118e-02</td>
</tr>
<tr>
<td>opt-$L_3$</td>
<td>3.816e-01</td>
<td>1.303e-01</td>
<td>1.145e-02</td>
</tr>
<tr>
<td>opt-$L_4$</td>
<td>9.368e-02</td>
<td>9.766e-03</td>
<td>3.383e-05</td>
</tr>
<tr>
<td>multi</td>
<td>9.393e-02</td>
<td>9.829e-03</td>
<td>3.469e-05</td>
</tr>
<tr>
<td>GCV-multi</td>
<td>9.411e-02</td>
<td>8.870e-03</td>
<td>5.076e-05</td>
</tr>
<tr>
<td>GCV1D-multi</td>
<td>3.013e-01</td>
<td>6.990e-02</td>
<td>1.533e-03</td>
</tr>
</tbody>
</table>
jointly diagonalizable by the 2D DFT. Given the training images, we use the Gauss–Newton approach described in section 3.2 to compute regularization parameters corresponding to the approximate problem and the 2-norm. These parameters are denoted as $\hat{\lambda}$, and reconstructions of images from the validation set corresponding to solving the original multi-Tikhonov problem with $\hat{\lambda}$ are denoted opt-multi-$\hat{\lambda}$. Results using optimal regularization parameters for the original problem are denoted opt-multi (this corresponds to figure 7). For comparison, we also provide results for multi-$\lambda_{GCV}$, which corresponds to solving the original multi-

**Figure 8.** Absolute error images, in inverted colormap so that white corresponds to small errors, for one of the validation images in example 2. Images on the left correspond to optimal parameters that were computed using training data using different $\rho$ functions. The last column of images contains absolute error images for reconstructions using GCV-selected parameters.
Tikhonov problem where parameters $\lambda_{GCV}$ were computed using GCV on the approximate problem. Boxplots in figure 9 provide statistics regarding the RREs for the validation images. We observe that reconstruction errors corresponding to learned parameters for the approximate problem are comparable to reconstruction errors corresponding to learned parameters for the original problem. Also, multi-$\lambda_{GCV}$ exhibits a slightly smaller interquartile range and smaller outliers, since GCV parameters are computed separately for each image in the validation set and can tailor the parameters to the individual characteristics of each validation image.

Example 3. We investigate the new S-Tik regularizing filter proposed in section 3.3 for image deblurring. The PSF, shown in figure 10(a), is $256 \times 256$ and separable. Thus $A$ can be represented using a Kronecker product and the SVD can be obtained [19]. We assume reflexive boundary conditions. For general-form and multi-parameter Tikhonov, obtaining a GSVD is not possible, so we consider a S-Tik filter where the subspace vectors are given by the eigenvalues of $L_4$, $L_2$, $L_3$, and $L_4$ from example 2. Images of these vectors are provided in figure 10(b).

A similar set-up as described in example 2 was used to generate training images, and the described learning methods were used to compute optimal S-Tik regularization parameters,
\( \lambda_{\text{STik}} \). For comparison purposes, we also used the training images to compute the opt-error-SVD filter. The first 5000 computed filter factors are provided in figure 11, in intervals of 5. It is worth noting that the computed filter factors are similar, but only 4 regularization parameters need to be computed for optimal S-Tik, whereas 65536 parameters need to be computed for optimal error-SVD. Furthermore, the optimal S-Tik filter can naturally enforce constraints such as non-negativity of the filter factors, by definition of the subspace.

The computed optimal S-Tik and opt-error-SVD filters were then used to reconstruct the blurred images in the validation set, and RREs are provided in table 2 for both the training and validation sets. We remark that the training set can be used to obtain uncertainty estimates for the reconstructions. For the training set, the average reconstruction error for the optimal error filter is smaller than that of the optimal S-Tik as expected, but the opposite is true for the validation data, which can be attributed to bias or overfitting to the training set. We also report the average error with standard deviation for GCV S-Tik, which corresponds to using the multi-parameter GCV function (4.4) for the S-Tik filter for each validation image separately. These results are similar to those presented in example 2, but it is worthwhile to mention the computational trade-off. In general, using GCV on each individual image can produce superior results but can be costly for problems where very large sets of images must be reconstructed in real time. For these problems, finding one set of parameters that will work for a variety of images will have a significant cost advantage and can produce results that are nearly as accurate as those obtained by individual GCV results.

Lastly, Pareto curves in figure 12 illustrate that good parameters for the optimal S-Tik filter can be obtained with fewer training images than the optimal error filter. For various
numbers of training images, we provide the median, 25th and 75th percentiles of the reconstruction errors for the set of validation images.

5. Conclusions

In this paper, we consider general-form and multi-parameter Tikhonov regularization. We describe a learning approach that uses training data to compute regularization parameters by formulating the problem as an empirical Bayes risk minimization problem and using efficient optimization schemes to compute regularization parameters that are optimal on average for the training set. In addition, we propose a new class of S-Tikhonov filters that computes a multi-parameter Tikhonov solution, with similar reconstruction quality as optimal error filters, but that requires a smaller training set and allows flexibility and novelty in the choice of regularization matrices. Numerical results illustrate that GSVD filtered solutions and S-Tikhonov filters can result in similar or better performance than optimal SVD filtered solution, with less training data. In addition, learned parameters for multi-parameter Tikhonov can provide results comparable to GCV, with less computational cost for a large data set.

References


[27] Liu J et al 2012 Morphology enabled dipole inversion for quantitative susceptibility mapping using structural consistency between the magnitude image and the susceptibility map Neuroimage 59 2560–8


[32] Peyré G and Nadif M 2011 Learning analysis sparsity priors Sampta’11