IOP Publishing Inverse Problems

Inverse Problems **39** (2023) 025004 (34pp) https://doi.org/10.1088/1361-6420/acad22

Parameter choices for sparse regularization with the ℓ_1 norm*

Qianru Liu[1], Rui Wang[1], Yuesheng Xu[2,] and Mingsong Yan[2]**

[1] School of Mathematics, Jilin University, Changchun 130012, People's Republic of China

[2] Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, United States of America

E-mail: y1xu@odu.edu

Received 14 May 2022; revised 23 November 2022
Accepted for publication 20 December 2022
Published 3 January 2023

Abstract

We consider a regularization problem whose objective function consists of a convex fidelity term and a regularization term determined by the ℓ_1 norm composed with a linear transform. Empirical results show that the regularization with the ℓ_1 norm can promote sparsity of a regularized solution. The goal of this paper is to understand theoretically the effect of the regularization parameter on the sparsity of the regularized solutions. We establish a characterization of the sparsity under the transform matrix of the solution. When the objective function is block-separable or an error bound of the regularized solution to a known function is available, the resulting characterization can be taken as a regularization parameter choice strategy with which the regularization problem has a solution having a sparsity of a certain level. When the objective function is not block-separable, we propose an iterative algorithm which simultaneously determines the regularization parameter and its corresponding solution with a prescribed sparsity level. Moreover, we study choices of the regularization parameter so that the regularization term can alleviate the ill-posedness and promote sparsity of the resulting regularized solution. Numerical experiments demonstrate that the proposed algorithm is effective and efficient, and the choices of the regularization parameters can balance the sparsity of the regularized solution and its approximation to the minimizer of the fidelity function.

* Dedicated to Professor Charles A Micchelli on the occasion of his 80th birthday with friendship and esteem.

** Author to whom any correspondence should be addressed.

Supplementary material for this article is available online

Keywords: parameter choice strategy, regularization, sparsity

(Some figures may appear in colour only in the online journal)

1. Introduction

Many practical problems may be modeled as learning a function from a finite number of observed data points. Learning a function from a finite number of observed data is an ill-posed problem. Such a problem cannot be solved directly as its solution is strongly sensitive to input data which are inevitably corrupted with noise. The ill-posedness was treated by the classical Tikhonov regularization which adds a regularization term to a data fidelity term constructed from the original ill-posed problem such that the resulting optimization problem is much less sensitive to disturbances. The added regularization term composes of a Hilbert space norm of the solution and a positive regularization parameter λ which balances the noise suppression and the approximation error of the regularized solution. An estimate for the classical Tikhonov regularization expresses the regularization error in terms of a sum of the two terms: the approximation error proportional to λ plus the error (caused by noise) proportional to the reciprocal of λ. The parameter λ is then chosen to minimize the regularization error. For choices of the optimal regularization parameter, the readers are referred to [5, 46, 65].

Motivated by the big data nature of recent practical applications, sparse regularization in Banach spaces has attracted much attention in various fields, since a sparse representation for a learned function is essential to ease the computational burden for operations of the function as the amount of data increases. As a popular approach to achieve this, regularization in a Banach space with a sparsity promoting norm, such as the ℓ_1 norm, is widely used in statistics, machine learning, signal processing, image processing and medical imaging. In statistics, the lasso and its extensions [1, 61–63] apply an ℓ_1 penalty to linear regression. The lasso is also known in signal processing as basis pursuit [15] which aims at decomposing a signal into an optimal superposition of dictionary elements in the sense that the resulting representation has the smallest ℓ_1 norm of coefficients among all such decompositions. For the purpose of solving nonlinear ill-posed problems, regularization with a one-homogeneous and convex constraint, which take the ℓ_1 norm as a typical example, was proposed in [48]. Image restoration using the total variation (TV) norm for regularization [35, 43, 49] leads to searching an optimization solution in the Euclidean space with the ℓ_1 norm. Sparse learning models with the ℓ_1 norm, such as ℓ_1 Support vector machine (SVM) classification [38, 51, 55] and ℓ_1 SVM regression [7, 37, 55], have received increasing attention in machine learning. Motivated by the need of sparse learning algorithms, the notion of reproducing kernel Banach spaces (RKBSs) was introduced in [76] and further developed in [40, 56, 57, 72]. RKBSs with the ℓ_1 norm [40, 56, 57] have been proven successful in promoting sparsity in representations for learned functions.

There were two crucial issues related to the choice of the regularization parameter in a regularization problem in a Banach space. The first one involves the error analysis to which considerable amount of work (for example, [26, 41, 52]) has been devoted. In particular, for the regularization problem with a special fidelity term and the ℓ_p norm regularizer, a convergence rate of the regularized solutions has been derived in [26, 41] according to a noise level and a choice the regularization parameter. The second issue concerns how a choice of the regularization parameter balances the sparsity of the regularized solution and its approximation accuracy. Empirical results [37, 49, 57, 61, 74] showed that one can obtain a solution having sparsity of certain level under a given transform of the regularization problem by choosing

appropriate regularization parameter. There also exist some theoretical results [4, 33, 54, 64, 77] for choices of the parameters in some special cases. For several specific application models, there were attempts to understand how one can choose the regularization parameter so that the resulting learned function has sparsity of certain levels. For example, sparsity of the solution of the lasso regularized model was studied in [4], where the relation between the sparsity of the regularized solution and the regularization parameter was characterized. Recent studies on the degrees of freedom of the lasso regularized model [64, 77] provided an objectively guided choice of the regularization parameter in such a regularization problem through Stein's unbiased risk estimation (SURE) framework. For the ℓ_1 regularized logistic regression problem, the regularization parameter is given in [33] to ensure that the regularized solution has all components zero. A sufficient condition for vanishing of a coefficient in a solution representation of a regularized learning method with an ℓ_1 regularizer was presented in [54]. These theoretical results on the regularization parameter all depend on the learned solution and no practical choice strategy of the parameter was provided in these studies.

It remains to be understood from the theoretical viewpoint for a regularized learning problem with a general convex fidelity term how the choice of the regularization parameter balances sparsity of the learned solution and its approximation error. The aim of this paper is to reveal theoretically how the choice of the regularization parameter can alleviate the ill-posedness and promoting sparsity of a regularized solution. To this end, we need to first study the relation between the choice of the regularization parameter and the sparsity of the regularized solutions in a Banach space with the ℓ_1 norm. This issue has been considered in [71] for the case when the regularization term is the ℓ_0 'norm' composed with a linear transform. Since the regularization problem with the ℓ_0 'norm' has nice geometric interpretation even though it is non-convex, it leads to a geometric approach to understand the issue. Since the ℓ_1 norm regularization problem has less clear geometric meaning, the geometric approach introduced in [71] does not seem to be applicable directly. However, it provides us with useful insights of sparse solutions. Due to the convexity of the ℓ_1 norm, we instead approach this problem by appealing to tools available in convex analysis.

In the regularization problem to be studied in this paper, the objective function consists of a convex fidelity term and a regularization term determined by the ℓ_1 norm composed with a transform matrix. We first discuss the choices of the regularization parameters when the transform matrix that appears in the regularization term reduces to the identity matrix. We have paid special attention to the cases when fidelity terms have special structures such as additive separability or block separability. In such cases, we have established a complete characterization of the sparsity of the solution, which show how we can choose the regularization parameter so that the solution has certain levels of sparsity. For the case that the fidelity term is a general convex function, we also give a sparsity characterization of the solution. Although in this characterization the regularization parameter depends on the solution, we still observe from it how the choice of the regularization parameter influences the sparsity of the solution. We then consider the case when the transform matrix is not the identity and has an arbitrary rank. In such a case, by making use of the singular value decomposition (SVD) of the transform matrix, we transform the original minimization problem to an equivalent constrained optimization problem having a simple transform matrix which is a two block diagonal matrix with the diagonal blocks being an identity and a zero matrix. The equivalent constrained optimization problem is further reformulated as an unconstrained minimization problem by employing the indicator function of the constraint set. In this manner, we obtain a characterization of the sparsity under the transform of the regularized solution. Results obtained in this paper are applied to several practical examples. Moreover, we conduct numerical experiments, to test the obtained theoretical results, which show that the parameter choices provided by this study

can balance the sparsity of the solution of the regularization problem and its approximation accuracy.

Choosing a regularization parameter to balance the sparsity of the regularized solution and its approximation accuracy in a Banach space setting is a challenging issue. Unlike the counterpart in a Hilbert space setting where the parameter was chosen to balance two error terms (the approximation error and the error caused by noise) which have the same base quantity (dimension) [16, 23, 59, 60], the sparsity measure and the accuracy measure in a Banach space setting are not the same base quantity. This raises technical difficulties in balancing them from a theoretical standpoint. We attempt in this paper to understand theoretically the effect of the regularization parameter to the sparsity of the regularized solution and as well as to approximation errors caused by noise. We demonstrate our idea by considering the lasso regularized model. An error estimate for a solution of the model can be obtained by a general argument established in [26]. By combining the sparsity characterization of the regularized solution and the error estimate, we obtain a choice strategy of the regularization parameter that yields a sparse regularized solution with an error bound.

Major contributions made in this paper are that we provide an implementable regularization parameter choice strategy, which balances the sparsity of the corresponding regularized solution and its approximation error bound, for the regularization problem with a block separable fidelity term. Moreover, for the case that the fidelity term is not block separable, we present a characterization of the regularization parameter which leads to a sparse regularized solution of a prescribed level. Based on such a characterization, we develop an iterative scheme for determining simultaneously the parameter and the associated regularized solution with a prescribed sparsity level, which also leads to an implementable regularization parameter choice strategy that produces a sparse regularized solution with an approximation error bound.

We organize this paper in seven sections and an appendix. In section 2, we describe the regularization problem to be considered and review several examples of practical importance. We characterize in section 3 the relation between the regularization parameter and the sparsity level of the regularized solution in the case that the transform matrix is the identity. The resulting characterizations provide regularization parameter choice strategies ensuring that the regularized solution with this parameter has sparsity of a desired level. Section 4 is devoted to studying choices of the regularization parameter that guarantee desired sparsity levels under a transform of the regularized solution and an iterative scheme that determines simultaneously the parameter and the associated regularized solution with a prescribed sparsity level. In section 5, we discuss how the regularization parameter λ can be chosen to alleviate the ill-posedness and promoting sparsity of the regularized solutions by considering a lasso regularized model. In section 6, we present numerical experiments to demonstrate the effectiveness of the regularization parameter choice strategy and the iterative algorithm established in this paper. In section 7, we make conclusive remarks. In appendix, we include proofs of several technical lemmas.

2. Regularization with the ℓ_1 norm

In this section, we describe the regularization problem to be considered in this paper, and identify several optimization models of practical importance which can be formulated in this general form.

We begin with describing the regularization problem. For each $d \in \mathbb{N}$, let $\mathbb{N}_d := \{1, 2, \ldots, d\}$ and set $\mathbb{N}_0 := \emptyset$. For $\mathbf{x} := [x_j : j \in \mathbb{N}_d] \in \mathbb{R}^d$, we define its ℓ_1 norm by $\|\mathbf{x}\|_1 := \sum_{j\in\mathbb{N}_d} |x_j|$. For $m, n \in \mathbb{N}$, suppose that $\psi : \mathbb{R}^n \to \mathbb{R}_+ := [0, +\infty)$ is a convex function and $\mathbf{B}$ is an $m \times n$ real matrix. We consider the regularization problem

$$\min\{\psi(\mathbf{u}) + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^n\}, \tag{1}$$

where λ is a positive regularization parameter. The regularization problem (1) covers many application problems. We present several examples below.

The generalized lasso regularized model [1, 63] is a special case of the regularization problem (1). Specifically, let $p \in \mathbb{N}$ and $\|\cdot\|_2$ denote the standard Euclidean norm on $\mathbb{R}^p$. Suppose that $\mathbf{x} \in \mathbb{R}^p$ is a response vector, $\mathbf{A} \in \mathbb{R}^{p\times n}$ is a predictor matrix and $\mathbf{B}$ is an $m \times n$ real matrix. When the fidelity term ψ is chosen as

$$\psi(\mathbf{u}) := \frac{1}{2}\|\mathbf{A}\mathbf{u} - \mathbf{x}\|_2^2, \quad \mathbf{u} \in \mathbb{R}^n, \tag{2}$$

the regularization problem (1) reduces to the generalized lasso regularized model

$$\min\left\{\frac{1}{2}\|\mathbf{A}\mathbf{u} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^n\right\}. \tag{3}$$

The regularization problem (3) covers many important areas where different choices of matrices $\mathbf{A}$ and $\mathbf{B}$ are taken. As a special case, the lasso regularized model [61] has the form (3) with $\mathbf{B}$ being the identity matrix $\mathbf{I}_n$ of order n.

In signal or image denoising processes, the two positive integers p and n are equal and the matrix $\mathbf{A}$ is chosen as the identity matrix of order n. The transform matrix $\mathbf{B}$ is often chosen to reflect some believed structure or geometry in the signal or the image. For example, if $\mathbf{B}$ is chosen as the $(n-1)\times n$ first order difference matrix $\mathbf{D}^{(1)} := [d_{ij} : i \in \mathbb{N}_{n-1}, j \in \mathbb{N}_n]$ with $d_{ii} = -1$, $d_{i,i+1} = 1$ for $i \in \mathbb{N}_{n-1}$ and 0 otherwise, then problem (3) describes the one-dimensional fused lasso model [62], which is also called the one-dimensional total-variation denoising model [18]. If $\mathbf{B}$ is chosen as the two-dimensional difference matrix giving both the horizontal and vertical differences between pixels, then problem (3) coincides with the two-dimensional fused lasso model [62] or the Rudin-Osher-Fatemi (ROF) total-variation denoising model [35, 43, 49].

Another example that concerns the polynomial trend filtering is described below. For each $k \in \mathbb{N}$, let $\mathbf{D}^{(1,k)}$ denote the $(n-k-1)\times(n-k)$ first order difference matrix. The difference matrix of order $k+1$ is defined recursively by $\mathbf{D}^{(k+1)} := \mathbf{D}^{(1,k)}\mathbf{D}^{(k)}$, $k \in \mathbb{N}$. The polynomial trend filtering of order k has the form (3) with $\mathbf{A} := \mathbf{I}_n$ and $\mathbf{B} := \mathbf{D}^{(k+1)}$. In the special case that $k = 1$, problem (3) reduces to the linear trend filtering [31, 70]. The transform matrix $\mathbf{B}$ may also be chosen as a discrete wavelet transform [14, 20, 22, 44, 66], a framelet transform [36, 39], a discrete cosine transform [28, 58] or a discrete Fourier transform [24, 42, 73], depending on specific applications in signal or imaging processing. The resulting regularized model aims at representing a signal or an image as a sparse linear combination of certain basis functions.

Data in many applications often carry a group structure where they are partitioned into disjoint pieces. Structured sparsity approaches recently received considerable attention in statistics, machine learning and signal processing. A natural extension of the lasso regularized model is the group lasso regularized model [10, 11, 29, 74]. We now briefly review this model. For $d, n \in \mathbb{N}$ with $d \leqslant n$, we suppose that $\mathcal{S} := \{S_1, S_2, \ldots, S_d\}$ is a partition of $\mathbb{N}_n$ in the sense that $S_j \neq \emptyset$, for all $j \in \mathbb{N}_d$, $S_j \cap S_k = \emptyset$ if $j \neq k$, and $\cup_{j\in\mathbb{N}_d} S_j = \mathbb{N}_n$. For each $j \in \mathbb{N}_d$ we denote by n_j the cardinality of S_j and regard S_j as an *ordered set* in the natural order of the elements in $\mathbb{N}_n$. That is, $S_j := \{i(j)_1, \ldots, i(j)_{n_j}\}$, with $i(j)_l \in \mathbb{N}_n$, $l \in \mathbb{N}_{n_j}$ and $i(j)_1 < \ldots < i(j)_{n_j}$. Associated with $\mathcal{S}$, we decompose $\mathbf{u} := [u_k : k \in \mathbb{N}_n] \in \mathbb{R}^n$ into d sub-vectors by setting $\mathbf{u}_j := [u_{i(j)_1}, \ldots, u_{i(j)_{n_j}}] \in \mathbb{R}^{n_j}$, $j \in \mathbb{N}_d$. The group lasso regularized model is described as

$$\min\left\{\frac{1}{2}\|\mathbf{A}\mathbf{u}-\mathbf{x}\|_2^2+\lambda\sum_{j\in\mathbb{N}_d}\sqrt{n_j}\|\mathbf{u}_j\|_2 : \mathbf{u}\in\mathbb{R}^n\right\}. \tag{4}$$

The regularizer in problem (4) could be viewed as a group-wise ℓ_1 norm. If the partition $\mathcal{S} := \{S_1, S_2, \ldots, S_n\}$ is chosen with $S_j := \{j\}, j \in \mathbb{N}_n$, then model (4) reduces to the lasso regularized model. It is known [30, 74] that model (4) performs better than the lasso regularized model when the optimal variable has the group structure.

SVMs for both classification and regression with the ℓ_1 norm can be reformulated in the form (1). We first present the ℓ_1 SVM classification model [51, 55]. Given training data $D := \{(\mathbf{x}_j, y_j) : j \in \mathbb{N}_n\}$ composed of input points $X := \{\mathbf{x}_j : j \in \mathbb{N}_n\} \subset \mathbb{R}^d$ and output values $Y := \{y_j : j \in \mathbb{N}_n\} \subset \{1, -1\}$. A hyperplane determined by $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ is constructed to separate D into two groups for $y_j = 1$ and $y_j = -1$ separately. By introducing a loss function $L_D : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}_+$, $\boldsymbol{\alpha}, b$ are obtained by the ℓ_1 SVM classification model

$$\min\{L_D(\boldsymbol{\alpha}, b) + \lambda\|\boldsymbol{\alpha}\|_1 : \boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}\}. \tag{5}$$

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a given reproducing kernel. A commonly used loss function L_D in model (5) is the hinge loss function defined by

$$L_D(\boldsymbol{\alpha}, b) := \sum_{j\in\mathbb{N}_n}\max\Big\{1 - y_j\Big(\sum_{k\in\mathbb{N}_n}\alpha_k K(\mathbf{x}_k, \mathbf{x}_j) + b\Big), 0\Big\},\ \boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}.$$

We next rewrite model (5) in the form (1). We let $\mathbf{u} := \begin{bmatrix}\boldsymbol{\alpha}\\ b\end{bmatrix} \in \mathbb{R}^{n+1}$, define the kernel matrix $\mathbf{K} := [K(\mathbf{x}_j, \mathbf{x}_k) : j, k \in \mathbb{N}_n]$ and augment it to $\mathbf{K}' := [\mathbf{K}\ 1_n]$ with $1_n := [1, \ldots, 1]^\top \in \mathbb{R}^n$. We also define $\mathbf{Y} := \operatorname{diag}(y_j : j \in \mathbb{N}_n)$ and $\phi(\mathbf{z}) := \sum_{j\in\mathbb{N}_n}\max\{1 - z_j, 0\}$, for all $\mathbf{z} := [z_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$. Then by introducing the fidelity term $\psi(\mathbf{u}) := \phi(\mathbf{Y}\mathbf{K}'\mathbf{u})$, $\mathbf{u} \in \mathbb{R}^{n+1}$ and choosing $\mathbf{B} := [\mathbf{I}_n\ 0] \in \mathbb{R}^{n\times(n+1)}$, the ℓ_1 SVM classification model (5) can be rewritten in the form of (1).

Another popular choice of the loss function L_D in the ℓ_1 SVM classification model (5) is the squared loss function defined by

$$L_D(\boldsymbol{\alpha}, b) := \frac{1}{2}\sum_{j\in\mathbb{N}_n}\Big(\sum_{k\in\mathbb{N}_n}\alpha_k K(\mathbf{x}_k, \mathbf{x}_j) + b - y_j\Big)^2. \tag{6}$$

By setting $\mathbf{y} := [y_j : j \in \mathbb{N}_n]$, the ℓ_1 SVM classification model (5) with the squared loss function (6) can be identified as the form (1) with $\psi(\mathbf{u}) := \frac{1}{2}\|\mathbf{K}'\mathbf{u} - \mathbf{y}\|_2^2$, $\mathbf{u} \in \mathbb{R}^{n+1}$, and $\mathbf{B} := [\mathbf{I}_n\ 0] \in \mathbb{R}^{n\times(n+1)}$. We note that this model can also be identified in the form of the generalized lasso model (3) with $\mathbf{x} := \mathbf{y}$, $\mathbf{A} := \mathbf{K}'$ and $\mathbf{B} := [\mathbf{I}_n\ 0]$.

When L_D in the ℓ_1 SVM classification model (5) is chosen as the average logistic loss function $L_D(\boldsymbol{\alpha}, b) := \frac{1}{n}\sum_{j\in\mathbb{N}_n}\ln\left(1 + \exp\left(-y_j\left(\boldsymbol{\alpha}^\top\mathbf{x}_j + b\right)\right)\right)$, for $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, it is the ℓ_1 regularized logistic regression model. It can be written in the form (1) with the fidelity term $\psi(\mathbf{u}) := \phi(\mathbf{Y}\mathbf{X}'\mathbf{u})$, $\mathbf{u} := \begin{bmatrix}\boldsymbol{\alpha}\\ b\end{bmatrix} \in \mathbb{R}^{d+1}$, and matrix $\mathbf{B} := [\mathbf{I}_d\ 0] \in \mathbb{R}^{d\times(d+1)}$, where $\mathbf{X} := [\mathbf{x}_j : j \in \mathbb{N}_n]^\top$, $\mathbf{X}' := [\mathbf{X}\ 1_n]$ and $\phi(\mathbf{z}) := \frac{1}{n}\sum_{j\in\mathbb{N}_n}\ln(1 + \exp(-z_j))$, for all $\mathbf{z} := [z_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$.

We now turn to describing the ℓ_1 SVM regression model [7, 55] which aims at learning a function from the observed data $D := \{(\mathbf{x}_j, y_j) : j \in \mathbb{N}_n\} \subset \mathbb{R}^d \times \mathbb{R}$. Specifically, the ℓ_1 SVM regression model has the same form as for the classification model (5), with a different loss function L_D. A popular choice of L_D is the ϵ-insensitive loss function [68] in the form

$$L_D(\boldsymbol{\alpha}, b) := \sum_{j\in\mathbb{N}_n}\max\Big\{\Big|\sum_{k\in\mathbb{N}_n}\alpha_k K(\mathbf{x}_k, \mathbf{x}_j) + b - y_j\Big| - \epsilon, 0\Big\}, \tag{7}$$

where ϵ is a positive parameter and K is a given reproducing kernel on $\mathbb{R}^d$. We may rewrite the ℓ_1 SVM regression model in the form of (1). To this end, we define the vector $\mathbf{u}$, the kernel matrix $\mathbf{K}$ and its augmented matrix $\mathbf{K}'$ as in the classification model. Associated with the output data values $\mathbf{y} := [y_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$ and $\epsilon > 0$, we introduce the function $\phi_{\mathbf{y},\epsilon}(\mathbf{z}) := \sum_{j\in\mathbb{N}_n} \max\{|z_j - y_j| - \epsilon, 0\}$, for all $\mathbf{z} := [z_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$. In this notation, the ℓ_1 SVM regression model with the ϵ-insensitive loss function (7) can be rewritten in the form (1) with the fidelity term $\psi(\mathbf{u}) := \phi_{\mathbf{y},\epsilon}(\mathbf{K}'\mathbf{u})$, $\mathbf{u} \in \mathbb{R}^{n+1}$, and matrix $\mathbf{B} := [\mathbf{I}_n\, 0] \in \mathbb{R}^{n\times(n+1)}$. The squared loss function L_D defined by (6) with $\mathbf{y} := [y_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$ is often used in the ℓ_1 SVM regression model. In this case, the ℓ_1 SVM regression model is equivalent to the regularization problem (1) composed of the fidelity term $\psi(\mathbf{u}) := \frac{1}{2}\|\mathbf{K}'\mathbf{u} - \mathbf{y}\|_2^2$, $\mathbf{u} \in \mathbb{R}^{n+1}$, with $\mathbf{y} := [y_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$ and the transform matrix $\mathbf{B} := [\mathbf{I}_n\, 0] \in \mathbb{R}^{n\times(n+1)}$.

The regularization problem (1) also appears in regularized learning in RKBSs. In such spaces, the regularized learning problem is usually an infinite dimensional optimization problem. The remarkable representer theorem [19, 32, 50, 67, 69] reduces the solutions to finding coefficients of a finite number of elements in the space. In particular, the regularized learning model in the RKBS with the ℓ_1 norm [56, 57] can be formulated in the form (1). Specifically, suppose that $\{(\mathbf{x}_j, y_j) : j \in \mathbb{N}_n\} \subset \mathbb{R}^d \times \mathbb{R}$ are given with $\mathbf{y} := [y_j : j \in \mathbb{N}_n]$, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a given reproducing kernel and $\mathbf{K} := [K(\mathbf{x}_j, \mathbf{x}_k) : j, k \in \mathbb{N}_n]$ is the resulting kernel matrix. The regularized learning model in the RKBS with the ℓ_1 norm has the form of (1) with $\psi(\mathbf{u}) := \|\mathbf{K}\mathbf{u} - \mathbf{y}\|_2^2$ and $\mathbf{B} := \mathbf{I}_n$.

3. Parameter choices for sparsity of the regularized solutions

In this section and the one followed, we discuss choices of the regularization parameter so that a solution of the resulting regularization problem (1) has sparsity of a prescribed level. In this section we first consider the special case when $m = n$ and $\mathbf{B} := \mathbf{I}_n$. In this case, the regularization problem (1) has the special form

$$\min\{\psi(\mathbf{u}) + \lambda\|\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^n\}. \tag{8}$$

We postpone the general case to the next section.

As a preparation, we recall the definition of the level of sparsity for a vector in $\mathbb{R}^n$. For each $n \in \mathbb{N}$, we set $\mathbb{Z}_n := \{0, 1, \ldots, n-1\}$. A vector $\mathbf{x} \in \mathbb{R}^n$ is said to have sparsity of level $l \in \mathbb{Z}_{n+1}$ if it has exactly l nonzero components. To further characterize sparsity of vectors in $\mathbb{R}^n$, we make use of the sparsity partition of $\mathbb{R}^n$, introduced initially in [71]. For each $j \in \mathbb{N}_n$, we denote by $\mathbf{e}_j$ the unit vector with 1 for the jth component and 0 otherwise. The vectors $\mathbf{e}_j, j \in \mathbb{N}_n$, form the canonical basis for $\mathbb{R}^n$. Using these vectors, we define $n+1$ numbers of subsets of $\mathbb{R}^n$ by $\Omega_0 := \{0 \in \mathbb{R}^n\}$ and $\Omega_l := \left\{\sum_{j\in\mathbb{N}_l} u_{k_j}\mathbf{e}_{k_j} : u_{k_j} \in \mathbb{R}\setminus\{0\},\ \text{for } 1 \leqslant k_1 < k_2 < \cdots < k_l \leqslant n\right\}$, for $l \in \mathbb{N}_n$. It was shown in [71] that the sets $\Omega_l, l \in \mathbb{Z}_{n+1}$, are mutually disjoint and form a partition for $\mathbb{R}^n$, that is, $\mathbb{R}^n = \bigcup_{l\in\mathbb{Z}_{n+1}} \Omega_l$. For each $l \in \mathbb{Z}_{n+1}$, Ω_l is the set of all vectors in $\mathbb{R}^n$ having sparsity of level l. Our goal is to relate the choice of the regularization parameter λ with the set Ω_l to which a solution $\mathbf{u}$ of the regularization problem (8) belongs.

We will employ the notion of the subdifferential of a convex function on $\mathbb{R}^n$ for this study. The subdifferential of a real-valued convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^n$ is defined by $\partial f(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n : f(\mathbf{z}) \geqslant f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} - \mathbf{x}\rangle,\ \text{for all } \mathbf{z} \in \mathbb{R}^n\}$. Suppose that f and g are two real-valued convex functions on $\mathbb{R}^n$. It is known [75] that if g is continuous on $\mathbb{R}^n$ then $\partial(f+g)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$.

We are now ready to characterize the sparsity of a solution of the regularization problem (8). We start with the case that the fidelity term ψ is block separable. For this purpose,

we let $\mathcal{S} := \{S_1, S_2, \ldots, S_d\}$ be a partition of $\mathbb{N}_n$, where for each $j \in \mathbb{N}_d$, we assume $S_j := \{i(j)_1, i(j)_2, \ldots, i(j)_{n_j}\}$, with $i(j)_l \in \mathbb{N}_n$, $l \in \mathbb{N}_{n_j}$ and $i(j)_1 < i(j)_2 < \ldots < i(j)_{n_j}$. For each $\mathbf{u} \in \mathbb{R}^n$, set $\mathbf{u}_j := [u_{i(j)_1}, u_{i(j)_2}, \ldots, u_{i(j)_{n_j}}]$ for all $j \in \mathbb{N}_d$. A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is called $\mathcal{S}$-block separable if there exist functions $\psi_j : \mathbb{R}^{n_j} \to \mathbb{R}$, $j \in \mathbb{N}_d$, such that

$$\psi(\mathbf{u}) = \sum_{j \in \mathbb{N}_d} \psi_j(\mathbf{u}_j), \text{ for all } \mathbf{u} \in \mathbb{R}^n. \tag{9}$$

A high-dimensional optimization problem having a block separable objective function can be reduced to several disjoint optimization problems with lower dimensionalities. By virtue of the block separability of ψ and the norm function $\|\cdot\|_1$, the regularization problem (8) can be reduced to the following lower dimensional regularization problems

$$\min\left\{\psi_j(\mathbf{u}_j) + \lambda\|\mathbf{u}_j\|_1 : \mathbf{u}_j \in \mathbb{R}^{n_j}\right\}, \ j \in \mathbb{N}_d. \tag{10}$$

We define the level of block sparsity for a vector in $\mathbb{R}^n$. We say that a vector $\mathbf{x} \in \mathbb{R}^n$ has $\mathcal{S}$-block sparsity of level $l \in \mathbb{Z}_{d+1}$ if $\mathbf{x}$ has exactly l number of nonzero sub-vectors.

We next give a choice of the parameter for the case when problem (8) has a most sparse solution without assuming ψ being block separable.

Lemma 3.1. *Suppose that ψ is a convex function on $\mathbb{R}^n$. Then the regularization problem (8) with $\lambda > 0$ has a solution $\mathbf{u}^* = 0$ if and only if $\lambda \geqslant \min\{\|\mathbf{y}\|_\infty : \mathbf{y} \in \partial\psi(0)\}$.*

Proof. According to the Fermat rule [75], vector $\mathbf{u}^* = 0$ is a solution of problem (8) if and only if $0 \in \partial(\psi + \lambda\|\cdot\|_1)(0)$, which by the continuity of the ℓ_1 norm is equivalent to $0 \in \partial\psi(0) + \lambda\partial\|\cdot\|_1(0)$. Hence, there exists $\mathbf{y} \in \partial\psi(0)$ such that $-\mathbf{y} \in \lambda\partial\|\cdot\|_1(0)$. Noting that $\partial\|\cdot\|_1(0) = \{\mathbf{y} \in \mathbb{R}^n : |y_j| \leqslant 1, j \in \mathbb{N}_n\}$, we rewrite $-\mathbf{y} \in \lambda\partial\|\cdot\|_1(0)$ as $\lambda \geqslant \|\mathbf{y}\|_\infty$. Thus, $\mathbf{u}^* = 0$ is a solution of (8) if and only if there exists $\mathbf{y} \in \partial\psi(0)$ such that $\lambda \geqslant \|\mathbf{y}\|_\infty$. It is clear that the latter is equivalent to $\lambda \geqslant \min\{\|\mathbf{y}\|_\infty : \mathbf{y} \in \partial\psi(0)\}$. □

With the help of lemma 3.1, we present a parameter choice so that a solution of problem (8), with ψ having the form (9), has block sparsity of a prescribed level.

Theorem 3.2. *Suppose that ψ_j, $j \in \mathbb{N}_d$, are convex functions on $\mathbb{R}^{n_j}$ and ψ is an $\mathcal{S}$-block separable function having the form (9). Then the regularization problem (8) with $\lambda > 0$ has a solution having the $\mathcal{S}$-block sparsity of level $l' \leqslant l$ for some $l \in \mathbb{Z}_{d+1}$ if and only if there exist distinct $k_i \in \mathbb{N}_d$, $i \in \mathbb{N}_l$, such that*

$$\lambda \geqslant \min\left\{\|\mathbf{y}\|_\infty : \mathbf{y} \in \partial\psi_j(0)\right\}, \text{ for all } j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{11}$$

In particular, if ψ_j, $j \in \mathbb{N}_d$, are differentiable, then condition (11) reduces to

$$\lambda \geqslant \|\nabla\psi_j(0)\|_\infty, \ \text{ for all } \ j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{12}$$

Proof. If $l = 0$, we give a choice of parameter λ so that problem (8) has the solution $\mathbf{u}^* = 0$. Note that $\mathbf{u}^* = 0$ is a solution of (8) if and only if for each $j \in \mathbb{N}_d$, $\mathbf{u}_j^* = 0$ is a solution of problem (10). Lemma 3.1 ensures that the latter holds if and only if $\lambda \geqslant \min\left\{\|\mathbf{y}\|_\infty : \mathbf{y} \in \partial\psi_j(0)\right\}$ for all $j \in \mathbb{N}_d$, which coincides with (11) with $l = 0$.

We next prove this theorem for the case that $l \neq 0$. We suppose that $\mathbf{u}^*$ as a solution of problem (8) has the $\mathcal{S}$-block sparsity of level $l' \leqslant l$ and $\mathbf{u}_j^*, j \in \mathbb{N}_d$, are its sub-vectors. That is, there exist distinct integers k_i, $i \in \mathbb{N}_{l'}$, in $\mathbb{N}_d$ such that $\mathbf{u}_j^* = 0$ for all $j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_{l'}\}$. Hence, problem (10) with $j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_{l'}\}$ has the trivial solution $\mathbf{u}_j^* = 0$. Again by lemma 3.1, we obtain that $\lambda \geqslant \min\left\{\|\mathbf{y}\|_\infty : \mathbf{y} \in \partial\psi_j(0)\right\}$, for all $j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_{l'}\}$. By choosing distinct integers $k_i \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_{l'}\}$, $i = l' + 1, \ldots, l$, we see that (11) follows. Conversely, suppose that there exist distinct $k_i \in \mathbb{N}_d$, $i \in \mathbb{N}_l$, such that (11) holds. By lemma 3.1, for each

$j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_l\}$, $\mathbf{u}_j^* = 0$ is a solution of problem (10). To construct a solution of problem (8), we choose for each $i \in \mathbb{N}_l$ a solution $\mathbf{u}_{k_i}^*$ of problem (10) with $j := k_i$. Let $\mathbf{u}^*$ be the vector in $\mathbb{R}^n$ with the sub-vectors $\mathbf{u}_j, j \in \mathbb{N}_d$, being defined above. It is clear that $\mathbf{u}^*$ is a solution for problem (8) and its level of $\mathcal{S}$-block sparsity is not more than l.

If $\boldsymbol{\psi}_j, j \in \mathbb{N}_d$, are differentiable, then their subdifferential at zero are the singleton $\nabla \boldsymbol{\psi}_j(0)$. This together with inequalities (11) leads to the desired inequalities (12). □

Theorem 3.2 reveals the relation between sparsity of a solution of problem (8) and a choice of parameter λ, when $\boldsymbol{\psi}$ is block separable. The choice of the parameter depends on the subdifferentials or the gradients of the functions $\boldsymbol{\psi}_j, j \in \mathbb{N}_d$.

As a special case, we consider problem (8) with $\boldsymbol{\psi}$ being additively separable. That is, there exist n univariate functions $\psi_j, j \in \mathbb{N}_n$, on $\mathbb{R}$ such that

$$\boldsymbol{\psi}(\mathbf{u}) := \sum_{j \in \mathbb{N}_n} \psi_j(u_j), \text{ for all } \mathbf{u} := [u_j : j \in \mathbb{N}_n] \in \mathbb{R}^n. \tag{13}$$

It is clear that an additively separable function $\boldsymbol{\psi}$ with the form (13) is $\mathcal{S}$-block separable with $\mathcal{S}$ being the nature partition of $\mathbb{N}_n$. That is, $S_j := \{j\}$. A parameter choice for this special case can be obtained directly from theorem 3.2. It is known [75] that for a convex function $\psi : \mathbb{R} \to \mathbb{R}$, both of its left derivative ψ'_- and its right derivative ψ'_+ exist at any $u \in \mathbb{R}$. Moreover, $\partial\psi(u) = \left[\psi'_-(u), \psi'_+(u)\right]$, for all $u \in \mathbb{R}$. For each $j \in \mathbb{N}_n$, let $\psi'_{j,-}$ and $\psi'_{j,+}$ denote the left and right derivatives of ψ_j, respectively.

Corollary 3.3. *Suppose that ψ_j, $j \in \mathbb{N}_n$, are convex functions on $\mathbb{R}$ and $\boldsymbol{\psi}$ has the form (13). Then problem (8) with $\lambda > 0$ has a solution having sparsity of level $l' \leqslant l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exist distinct $k_i \in \mathbb{N}_n$, $i \in \mathbb{N}_l$, such that $\lambda \geqslant \max\left\{0, \psi'_{j,-}(0), -\psi'_{j,+}(0)\right\}$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$. If ψ_j, $j \in \mathbb{N}_n$, are differentiable, then above condition reduces to $\lambda \geqslant |\psi'_j(0)|$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$.*

Proof. Note that $\boldsymbol{\psi}$ with the form (13) is $\mathcal{S}$-block separable with $\mathcal{S}$ being the nature partition of $\mathbb{N}_n$. Theorem 3.2 ensures that problem (8) with $\lambda > 0$ has a solution having sparsity of level $l' \leqslant l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exist distinct $k_i \in \mathbb{N}_n$, $i \in \mathbb{N}_l$, such that $\lambda \geqslant \min\{|y| : y \in \partial\psi_j(0)\}$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$. Noting that $\partial\psi_j(0) = [\psi'_{j,-}(0), \psi'_{j,+}(0)]$, for all $j \in \mathbb{N}_n$, the latter is equivalent to $\lambda \geqslant \max\left\{0, \psi'_{j,-}(0), -\psi'_{j,+}(0)\right\}$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$. If $\psi_j, j \in \mathbb{N}_n$, are differentiable, then above inequalities reduces to $\lambda \geqslant |\psi'_j(0)|$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$. □

In the following, we consider the lasso regularized model and discuss when the fidelity term $\boldsymbol{\psi}$ defined by (2) is block separable. Throughout this paper, we denote the jth column of a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ by $\mathbf{M}_j$. Associated with the partition $\mathcal{S}$, we decompose a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ into d sub-matrices by setting $\mathbf{M}_{(j)} := [\mathbf{M}_k : k \in S_j] \in \mathbb{R}^{m \times n_j}$ for all $j \in \mathbb{N}_d$. The next lemma provides a sufficient and necessary condition ensuring the block separability of the fidelity term $\boldsymbol{\psi}$ defined by (2).

Lemma 3.4. *Suppose that $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{p \times n}$. Then the function $\boldsymbol{\psi}$ defined by (2) is $\mathcal{S}$-block separable if and only if there holds*

$$(\mathbf{A}_{(j)})^\top \mathbf{A}_{(k)} = 0, \text{ for all } j, k \in \mathbb{N}_d \text{ and } j \neq k. \tag{14}$$

Proof. According to the definition (2) of $\boldsymbol{\psi}$, we have that $\boldsymbol{\psi}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top \mathbf{A}^\top \mathbf{A}\mathbf{u} - \mathbf{x}^\top \mathbf{A}\mathbf{u} + \frac{1}{2}\mathbf{x}^\top \mathbf{x}$, for all $\mathbf{u} \in \mathbb{R}^n$. It follows from the decomposition of $\mathbf{A}$ and that of each vector $\mathbf{u}$ in

$\mathbb{R}^n$ with respect to $\mathcal{S}$ that $\mathbf{A}\mathbf{u} = \sum_{j\in\mathbb{N}_d} \mathbf{A}_{(j)}\mathbf{u}_j$, for all $\mathbf{u}\in\mathbb{R}^n$. Substituting the above equation into the representation of ψ, we obtain that

$$\psi(\mathbf{u}) = \frac{1}{2}\sum_{j\in\mathbb{N}_d}\sum_{k\in\mathbb{N}_d} \mathbf{u}_j^\top (\mathbf{A}_{(j)})^\top \mathbf{A}_{(k)}\mathbf{u}_k - \sum_{j\in\mathbb{N}_d} \mathbf{x}^\top \mathbf{A}_{(j)}\mathbf{u}_j + \frac{1}{2}\mathbf{x}^\top\mathbf{x}, \text{ for all } \mathbf{u}\in\mathbb{R}^n. \tag{15}$$

Clearly, the last two terms in the right hand side of equation (15) are both $\mathcal{S}$-block separable. Hence, ψ is $\mathcal{S}$-block separable if and only if the first term is $\mathcal{S}$-block separable. The latter one is equivalent to that condition (14) holds. □

We now apply theorem 3.2 to the lasso regularized model when the matrix $\mathbf{A}$ satisfies condition (14).

Corollary 3.5. *Suppose that $\mathbf{x}\in\mathbb{R}^p$, $\mathbf{A}\in\mathbb{R}^{p\times n}$ and condition (14) holds. Then the lasso regularized model with $\lambda>0$ has a solution having the $\mathcal{S}$-block sparsity of level $l'\leqslant l$ for some $l\in\mathbb{Z}_{d+1}$ if and only if there exist distinct $k_i\in\mathbb{N}_d$, $i\in\mathbb{N}_l$, such that $\lambda \geqslant \left\|(\mathbf{A}_{(j)})^\top\mathbf{x}\right\|_\infty$, for all $j\in\mathbb{N}_d\setminus\{k_i : i\in\mathbb{N}_l\}$.*

Proof. Since condition (14) holds, lemma 3.4 ensures that the fidelity term ψ involved in the lasso regularized model is $\mathcal{S}$-block separable. Substituting condition (14) into equation (15), ψ can be represented in the form (9) with ψ_j, $j\in\mathbb{N}_d$, being defined by

$$\psi_j(\mathbf{u}_j) := \frac{1}{2}\|\mathbf{A}_{(j)}\mathbf{u}_j\|_2^2 - \mathbf{x}^\top\mathbf{A}_{(j)}\mathbf{u}_j + \frac{1}{2d}\mathbf{x}^\top\mathbf{x}, \text{ for all } \mathbf{u}_j\in\mathbb{R}^{n_j} \text{ and all } j\in\mathbb{N}_d. \tag{16}$$

Thus, we conclude by theorem 3.2 that the lasso regularized model has a solution having the $\mathcal{S}$-block sparsity of level $l'\leqslant l$ for some $l\in\mathbb{Z}_{d+1}$ if and only if there exist distinct integers k_i, $i\in\mathbb{N}_l$, in $\mathbb{N}_d$ such that inequality (12) holds. Substituting $\nabla\psi_j(0) = -(\mathbf{A}_{(j)})^\top\mathbf{x}$ for all $j\in\mathbb{N}_d$ into (12) leads directly to the desired inequalities. □

In signal or imaging processing, the matrix $\mathbf{A}$ involved in the lasso regularized model is often chosen as an orthogonal matrix, such as an orthogonal wavelet transform. We note that an orthogonal matrix is a special matrix satisfying condition (14) for any partition $\mathcal{S}$ of the index set $\mathbb{N}_n$. Especially, condition (14) holds for the nature partition $\mathcal{S}$ of $\mathbb{N}_n$. In this case, corollary 3.5 ensures that the lasso regularized model with $\lambda>0$ has a solution having sparsity of level $l'\leqslant l$ for some $l\in\mathbb{Z}_{n+1}$ if and only if there exist distinct $k_i\in\mathbb{N}_n$, $i\in\mathbb{N}_l$, such that $\lambda\geqslant\left|(\mathbf{A}_j)^\top\mathbf{x}\right|$, for all $j\in\mathbb{N}_n\setminus\{k_i : i\in\mathbb{N}_l\}$.

We next consider the group lasso regularized model (4) which is designed to obtain the block sparsity of the solutions. To describe a choice of the parameter for this regularization problem, we also assume that condition (14) holds. Then by lemma 3.4, the fidelity term in this problem is $\mathcal{S}$-block separable and has the form (9) with ψ_j, $j\in\mathbb{N}_d$, being defined by (16). It is obvious that the regularizer of the group lasso regularized model (4) is also $\mathcal{S}$-block separable. Therefore, it can be reduced to d lower dimensional regularization problems

$$\min\left\{\psi_j(\mathbf{u}_j) + \lambda\sqrt{n_j}\|\mathbf{u}_j\|_2 : \mathbf{u}_j\in\mathbb{R}^{n_j}\right\},\ j\in\mathbb{N}_d. \tag{17}$$

Through characterizing the sparsity of the solutions of problem (17), we obtain the following parameter choice with which the group lasso regularized model (4) has a solution having block sparsity of a prescribed level.

Theorem 3.6. *Suppose that $\mathbf{x}\in\mathbb{R}^p$, $\mathbf{A}\in\mathbb{R}^{p\times n}$ and condition (14) holds. Then problem (4) with $\lambda>0$ has a solution having the $\mathcal{S}$-block sparsity of level $l'\leqslant l$ for some $l\in\mathbb{Z}_{d+1}$ if and only if there exist distinct $k_i\in\mathbb{N}_d$, $i\in\mathbb{N}_l$, such that $\lambda\geqslant\left\|(\mathbf{A}_{(j)})^\top\mathbf{x}\right\|_2/\sqrt{n_j}$, for all $j\in\mathbb{N}_d\setminus\{k_i : i\in\mathbb{N}_l\}$.*

Proof. Condition (14) ensures that the fidelity term in problem (4) is $\mathcal{S}$-block separable and thus, $\mathbf{u}^*$ is a solution of (4) if and only if for each $j \in \mathbb{N}_d$, $\mathbf{u}_j^*$ is a solution of (17). It suffices to show that for each $j \in \mathbb{N}_d$, problem (17) has a solution $\mathbf{u}_j^* = 0$ if and only if $\lambda \geqslant \|(\mathbf{A}_{(j)})^\top \mathbf{x}\|_2/\sqrt{n_j}$. It follows from the Fermat rule and the differentiability of ψ_j that $\mathbf{u}_j^* = 0$ is a solution of (17) if and only if $0 \in \nabla\psi_j(0) + \lambda\sqrt{n_j}\partial\|\cdot\|_2(0)$. Since $\nabla\psi_j(0) = -(\mathbf{A}_{(j)})^\top \mathbf{x}$ and $\partial\|\cdot\|_2(0) = \{\mathbf{y} \in \mathbb{R}^{n_j} : \|\mathbf{y}\|_2 \leqslant 1\}$, the above inclusion relation is equivalent to inequality $\lambda \geqslant \left\|(\mathbf{A}_{(j)})^\top \mathbf{x}\right\|_2/\sqrt{n_j}$. Consequently, we conclude that $\mathbf{u}_j^* = 0$ is a solution of (17) if and only if the above inequality holds. By arguments similar to those used in the proof of theorem 3.2 and by employing inequality $\lambda \geqslant \left\|(\mathbf{A}_{(j)})^\top \mathbf{x}\right\|_2/\sqrt{n_j}$, we get the desired conclusion of this theorem. □

Many applications can be modeled as in the form (8) with ψ neither additively separable nor block separable. The next theorem concerns a sparsity characterization of the solution of problem (8) when ψ is a general convex on $\mathbb{R}^n$. For each $j \in \mathbb{N}_n$, we denote by ψ_j' the partial derivative of ψ with respect to the jth variable.

Theorem 3.7. *Suppose that ψ is a convex function on $\mathbb{R}^n$. Then problem (8) with $\lambda > 0$ has a solution $\mathbf{u}^* = \sum_{i\in\mathbb{N}_l} u_{k_i}^* \mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exists $\mathbf{y} := [y_j : j \in \mathbb{N}_n] \in \partial\psi(\mathbf{u}^*)$ such that*

$$\lambda = -y_{k_i}\operatorname{sign}(u_{k_i}^*),\ i \in \mathbb{N}_l \ \text{ and } \ \lambda \geqslant |y_j|,\ j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{18}$$

In particular, if ψ is a differentiable, then condition (18) is equivalent to

$$\lambda = -\psi_{k_i}'(\mathbf{u}^*)\operatorname{sign}(u_{k_i}^*),\ i \in \mathbb{N}_l \ \text{ and } \ \lambda \geqslant |\psi_j'(\mathbf{u}^*)|,\ j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{19}$$

Proof. By using the Fermat rule and the continuity of the ℓ_1 norm, we conclude that $\mathbf{u}^*$ is a solution of problem (8) if and only if $0 \in \partial\psi(\mathbf{u}^*) + \lambda\partial\|\cdot\|_1(\mathbf{u}^*)$. Equivalently, there exists $\mathbf{y} := [y_j : j \in \mathbb{N}_n] \in \partial\psi(\mathbf{u}^*)$ such that $-\mathbf{y} \in \lambda\partial\|\cdot\|_1(\mathbf{u}^*)$. Noting that $\mathbf{u}^* = \sum_{i\in\mathbb{N}_l} u_{k_i}^* \mathbf{e}_{k_i}$ with $u_{k_i}^* \in \mathbb{R}\setminus\{0\}$, $i \in \mathbb{N}_l$, we obtain that $\partial\|\cdot\|_1(\mathbf{u}^*) = \left\{\mathbf{z} \in \mathbb{R}^n : z_{k_i} = \operatorname{sign}(u_{k_i}^*), i \in \mathbb{N}_l \text{ and } |z_j| \leqslant 1, j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}\right\}$. By using the above equation, we rewrite inclusion relation $-\mathbf{y} \in \lambda\partial\|\cdot\|_1(\mathbf{u}^*)$ as (18).

If ψ is differentiable, then the subdifferential of ψ at $\mathbf{u}^*$ is the singleton $\nabla\psi(\mathbf{u}^*)$. Substituting $y_j = \psi_j'(\mathbf{u}^*), j \in \mathbb{N}_n$, into (18) leads directly to (19). □

We now apply theorem 3.7 to the lasso regularized model. Here, the fidelity term ψ defined by (2) is differentiable but neither additively separable nor block separable.

Corollary 3.8. *Suppose that $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{p\times n}$ are given. Then the lasso regularized model with $\lambda > 0$ has a solution $\mathbf{u}^* = \sum_{i\in\mathbb{N}_l} u_{k_i}^* \mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there hold $\lambda = (\mathbf{A}_{k_i})^\top(\mathbf{x} - \mathbf{A}\mathbf{u}^*)\operatorname{sign}(u_{k_i}^*)$, for all $i \in \mathbb{N}_l$ and $\lambda \geqslant \left|(\mathbf{A}_j)^\top(\mathbf{A}\mathbf{u}^* - \mathbf{x})\right|$, for all $j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}$.*

Proof. Note that the gradient of ψ at $\mathbf{u}^*$ has the form $\nabla\psi(\mathbf{u}^*) = \mathbf{A}^\top(\mathbf{A}\mathbf{u}^* - \mathbf{x})$. That is, $\psi_j'(\mathbf{u}^*) = (\mathbf{A}_j)^\top(\mathbf{A}\mathbf{u}^* - \mathbf{x})$, for all $j \in \mathbb{N}_n$. Theorem 3.7 ensures that $\mathbf{u}^* = \sum_{i\in\mathbb{N}_l} u_{k_i}^* \mathbf{e}_{k_i} \in \Omega_l$ is a solution of the lasso regularized model if and only if (19) holds. According to the representations of the partial derivatives of ψ, we conclude that the latter is equivalent to the desired results of this corollary. □

As a special case of corollary 3.8, the lasso regularized model has $\mathbf{u}^* = 0$ as a solution if and only if there holds $\lambda \geqslant \left\|\mathbf{A}^\top \mathbf{x}\right\|_\infty$. We remark that the special case of corollary 3.8 for the lasso regularized model has been established in [4].

When the fidelity term ψ has no special form such as (9) and (13), theorem 3.7 provides a characterization of the regularization parameter with which problem (8) has a solution with sparsity of a certain level. In fact, since condition (18) (or (19)) depends on the corresponding solution, the characterization stated in theorem 3.7 can not be used directly as a parameter choice strategy. Nevertheless, we can still observe from the characterization that the choice of the regularization parameter can influence the sparsity of the solution. The equalities and inequalities that the parameter λ needs to satisfy corresponds respectively the non-zero components and the zero components of the solution. As the number of the inequalities increases, the solution becomes more sparse. When the conditions only include the inequalities, the solution is the most sparse. This observation has motivated us to develop an iteration scheme for parameter choices so that a solution of the resulting regularization problem has sparsity of a prescribed level. We will present this algorithm in section 4 in the setting where $\mathbf{B}$ emerging in problem (1) is a general matrix.

4. Parameter choices for sparsity of transformed solutions

In this section, we continue our investigation about what choices of the regularization parameter lead to sparsity under a general transform matrix $\mathbf{B}$ for the solutions of problem (1). We first employ the SVD of matrix $\mathbf{B}$ to convert problem (1) to one with $\mathbf{B}$ being a *degenerated* identity (an identity matrix augmented by zero matrices). Based on this result, we characterize the regularization parameter for a sparse regularized solution of the resulting regularization problem by using the approach used in section 3 for the case when $\mathbf{B} := \mathbf{I}$. We then present special results for several specific learning models.

We now characterize the sparsity of a solution under a transform matrix $\mathbf{B}$ of problem (1). If $\mathbf{B}$ is an invertible square matrix, then by a simple change of variables problem (1) can be converted to problem (8). For a general matrix $\mathbf{B}$, we appeal to its pseudoinverse (Moore–Penrose inverse) [27]. To this end, we review the notion of the SVD of a matrix [27]. Suppose that $\mathbf{B}$ is a real $m \times n$ matrix with the rank r satisfying $0 < r \leqslant \min\{m,n\}$. It is well-known that $\mathbf{B}$ has the SVD as $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{\Lambda}$ is an $m \times n$ diagonal matrix with the singular values $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$ on the diagonal, and $\mathbf{V}$ is an $n \times n$ orthogonal matrix. A matrix, denoted by $\mathbf{M}^\dagger$, is called the pseudoinverse of $\mathbf{M}$ if it satisfies the four conditions: (a) $\mathbf{M}\mathbf{M}^\dagger\mathbf{M} = \mathbf{M}$, (b) $\mathbf{M}^\dagger\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger$, (c) $(\mathbf{M}\mathbf{M}^\dagger)^\top = \mathbf{M}\mathbf{M}^\dagger$, (d) $(\mathbf{M}^\dagger\mathbf{M})^\top = \mathbf{M}^\dagger\mathbf{M}$. The pseudoinverse is well-defined and unique for all matrices. It can be readily verified that the pseudoinverse $\mathbf{\Lambda}^\dagger$ of the $m \times n$ diagonal matrix $\mathbf{\Lambda}$ is the $n \times m$ diagonal matrix with the nonzero diagonal entries $\sigma_1^{-1}, \ldots, \sigma_r^{-1}$. Thus, the pseudoinverse $\mathbf{B}^\dagger$ of $\mathbf{B}$ can be represented by the SVD of $\mathbf{B}$ as $\mathbf{B}^\dagger = \mathbf{V}\mathbf{\Lambda}^\dagger\mathbf{U}^\top$.

In the next lemma, we consider inverting the linear system

$$\mathbf{B}\mathbf{u} = \mathbf{z}, \quad \text{for } \mathbf{z} \in \mathcal{R}(\mathbf{B}). \tag{20}$$

Here, $\mathcal{R}(\mathbf{B})$ denotes the range of $\mathbf{B}$. Note that solutions of system (20) may not be unique since if $\mathbf{u}'$ is a solution of (20), then $\mathbf{u} := \mathbf{u}' + \mathbf{u}_0$ is a solution of (20), for any $\mathbf{u}_0$ satisfying $\mathbf{B}\mathbf{u}_0 = 0$. It is known from [8] that by choosing $\mathbf{u}' := \mathbf{B}^\dagger\mathbf{z}$ as a particular solution of (20), the general solution of (20) has the form $\mathbf{u} = \mathbf{B}^\dagger\mathbf{z} + \mathbf{V}\begin{bmatrix} 0 \\ \mathbf{v} \end{bmatrix}$ for $0 \in \mathbb{R}^r$ and any $\mathbf{v} \in \mathbb{R}^{n-r}$. To convert problem (1) to an equivalent one, we give in the next lemma an alternative form of the general solution of (20), whose proof is included in appendix. Let $\widetilde{\mathbf{U}}_r \in \mathbb{R}^{m \times r}$ denote the matrix composed of the first r columns of $\mathbf{U}$. We introduce a diagonal matrix of order n by $\mathbf{\Lambda}' := \operatorname{diag}\left(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_r^{-1}, 1, \ldots, 1\right)$ and an $n \times (m+n-r)$ block diagonal matrix $\mathbf{U}' :=$

diag $\left(\widetilde{\mathbf{U}}_r^\top, \mathbf{I}_{n-r}\right)$. Using these matrices, we define an $n \times (m+n-r)$ matrix $\mathbf{B}' := \mathbf{V}\mathbf{\Lambda}'\mathbf{U}'$. In the following presentations, we always assume that $\mathbf{B}$ is a real $m \times n$ matrix with the SVD $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$.

Lemma 4.1. *If $\mathbf{z} \in \mathcal{R}(\mathbf{B})$, then the general solution of system (20) has the form*

$$\mathbf{u} = \mathbf{B}' \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix}, \ \text{for any} \ \ \mathbf{v} \in \mathbb{R}^{n-r}. \tag{21}$$

Moreover, for each solution $\boldsymbol{u}$, the vector $\boldsymbol{v}$ satisfying (21) is unique.

Lemma 4.1 allows us to introduce a mapping $\mathcal{B}$ from $\mathbb{R}^n$ to $\mathcal{R}(\mathbf{B}) \times \mathbb{R}^{n-r}$. Specifically, for each $\mathbf{u} \in \mathbb{R}^n$, we define $\mathcal{B}\mathbf{u} := \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix}$, where $\mathbf{z} := \mathbf{B}\mathbf{u}$ and $\mathbf{v} \in \mathbb{R}^{n-r}$ satisfies $\mathbf{u} = \mathbf{B}' \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix}$. Lemma 4.1 ensures that $\mathcal{B}$ is well-defined and satisfies $\mathbf{B}'\mathcal{B}\mathbf{u} = \mathbf{u}$, for all $\mathbf{u} \in \mathbb{R}^n$. Next lemma shows the bijectivity of $\mathcal{B}$ with a proof included in appendix.

Lemma 4.2. *The mapping $\mathcal{B}$ defined as above is bijective from $\mathbb{R}^n$ onto $\mathcal{R}(\mathbf{B}) \times \mathbb{R}^{n-r}$.*

We now reformulate problem (1) as an equivalent constrained regularization problem with $\mathbf{B}$ being a degenerated identity $\mathbf{I}' := [\mathbf{I}_m \ 0] \in \mathbb{R}^{m \times (m+n-r)}$, that is,

$$\min \left\{ \psi \circ \mathbf{B}'(\mathbf{w}) + \lambda \|\mathbf{I}'\mathbf{w}\|_1 : \mathbf{w} \in \mathcal{R}(\mathbf{B}) \times \mathbb{R}^{n-r} \right\}. \tag{22}$$

Note that the first m components of $\mathbf{w}$ is constrained to $\mathcal{R}(\mathbf{B})$. The proof of the following proposition is also included in appendix.

Proposition 4.3. *If $\mathcal{B}$ is defined as above, then $\mathbf{u}^*$ is a solution of the regularization problem (1) if and only if $\mathcal{B}\mathbf{u}^*$ is a solution of the regularization problem (22).*

We reformulate problem (22) as an equivalent unconstrained problem for the purpose of characterizing its sparse solutions. Set $\mathbb{M} := \mathcal{R}(\mathbf{B}) \times \mathbb{R}^{n-r}$ and denote by $\mathbb{M}^\perp$ its orthogonal complement. Let $\iota_{\mathbb{M}} : \mathbb{R}^{m+n-r} \to \mathbb{R} \cup \{+\infty\}$ denote the indicator function of $\mathbb{M}$, that is, $\iota_{\mathbb{M}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathbb{M}$, and $+\infty$ otherwise. Using the indicator function, the constrained problem (22) is rewritten as the equivalent unconstrained problem

$$\min \left\{ \psi \circ \mathbf{B}'(\mathbf{w}) + \iota_{\mathbb{M}}(\mathbf{w}) + \lambda \|\mathbf{I}'\mathbf{w}\|_1 : \mathbf{w} \in \mathbb{R}^{m+n-r} \right\}. \tag{23}$$

We present below a characterization of a solution of problem (23) having sparsity of a certain level. Let $\mathcal{N}(\mathbf{A})$ denote the null space of matrix $\mathbf{A}$. We suppose through out this section that ψ is a convex function on $\mathbb{R}^n$.

Proposition 4.4. *The problem (23) with $\lambda > 0$ has a solution $\mathbf{w}^* := \begin{bmatrix} \mathbf{z}^* \\ \mathbf{v}^* \end{bmatrix}$ with $\mathbf{z}^* := \sum_{i \in \mathbb{N}_l} z_{k_i}^* \mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{m+1}$ and distinct $k_i \in \mathbb{N}_n$, $i \in \mathbb{N}_l$ if and only if there exist $\mathbf{a} \in \partial\psi(\mathbf{B}'\mathbf{w}^*)$ and $\mathbf{b} := [b_j : j \in \mathbb{N}_m] \in \mathcal{N}(\mathbf{B}^\top)$ such that*

$$(\mathbf{B}'_j)^\top \mathbf{a} = 0, \ j \in \mathbb{N}_{m+n-r} \setminus \mathbb{N}_m, \ \ \lambda = -\left((\mathbf{B}'_{k_i})^\top \mathbf{a} + b_{k_i}\right) \operatorname{sign}(z_{k_i}^*), \ i \in \mathbb{N}_l, \tag{24}$$

$$\lambda \geqslant \left|(\mathbf{B}'_j)^\top \mathbf{a} + b_j\right|, \ j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{25}$$

Proof. According to the Fermat rule and the chain rule of the subdifferential, we get that $\mathbf{w}^* := \begin{bmatrix} \mathbf{z}^* \\ \mathbf{v}^* \end{bmatrix}$ is a solution of (23) if and only if

$$0 \in (\mathbf{B}')^\top \partial\psi(\mathbf{B}'\mathbf{w}^*) + \partial\iota_{\mathbb{M}}(\mathbf{w}^*) + \lambda(\mathbf{I}')^\top \partial\|\cdot\|_1(\mathbf{z}^*). \tag{26}$$

It is known that $\partial\iota_{\mathbb{M}}(\mathbf{w}) = \mathbb{M}^\perp$ for all $\mathbf{w} \in \mathbb{M}$, which together with $\mathbb{M}^\perp = (\mathcal{R}(\mathbf{B}))^\perp \times (\mathbb{R}^{n-r})^\perp = \mathcal{N}(\mathbf{B}^\top) \times \{0\}$ further leads to $\partial\iota_{\mathbb{M}}(\mathbf{w}) = \mathcal{N}(\mathbf{B}^\top) \times \{0\}$, for all $\mathbf{w} \in \mathbb{M}$. By

employing the above equation and noting that $\mathbf{w}^* \in \mathbb{M}$, the inclusion relation (26) is equivalent to the existence of $\mathbf{a} \in \partial\psi(\mathbf{B}'\mathbf{w}^*)$ and $\mathbf{b} \in \mathcal{N}(\mathbf{B}^\top)$ satisfying $-(\mathbf{B}')^\top\mathbf{a} - \begin{bmatrix}\mathbf{b}\\0\end{bmatrix} \in \lambda(\mathbf{I}')^\top\partial\|\cdot\|_1(\mathbf{z}^*)$. This inclusion relation together with

$$\partial\|\cdot\|_1(\mathbf{z}^*) = \left\{\mathbf{x}\in\mathbb{R}^m : x_{k_i} = \operatorname{sign}(z^*_{k_i}), i\in\mathbb{N}_l \text{ and } |x_j|\leqslant 1, j\in\mathbb{N}_m\setminus\{k_i : i\in\mathbb{N}_l\}\right\}$$

may be rewritten equivalently as (24) and (25). This proves the desired result. □

Combining propositions 4.3 and 4.4, we characterize a solution of the regularization problem (1) having sparsity of a certain level under the transform $\mathbf{B}$.

Theorem 4.5. *The problem (1) with $\lambda > 0$ has a solution $\mathbf{u}^* \in \mathbb{R}^n$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{m+1}$ if and only if there exist $\mathbf{a} \in \partial\psi(\mathbf{u}^*)$ and $\mathbf{b} \in \mathcal{N}(\mathbf{B}^\top)$ such that (24) and (25) hold. In particular, if $\operatorname{rank}(\mathbf{B}) = m$, then the conditions reduce to that there exists $\mathbf{a} \in \partial\psi(\mathbf{u}^*)$ such that*

$$(\mathbf{B}'_j)^\top\mathbf{a} = 0,\ j\in\mathbb{N}_n\setminus\mathbb{N}_m,\ \ \lambda = -(\mathbf{B}'_{k_i})^\top\mathbf{a}\operatorname{sign}(z^*_{k_i}),\ i\in\mathbb{N}_l, \tag{27}$$

$$\lambda \geqslant \left|(\mathbf{B}'_j)^\top\mathbf{a}\right|,\ j\in\mathbb{N}_m\setminus\{k_i : i\in\mathbb{N}_l\}. \tag{28}$$

Proof. By proposition 4.3 we conclude that $\mathbf{u}^* \in \mathbb{R}^n$ is a solution of (1) and $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ if and only if $\mathcal{B}\mathbf{u}^* := \begin{bmatrix}\mathbf{z}^*\\\mathbf{v}^*\end{bmatrix}$ with $\mathbf{z}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ is a solution of (23). Proposition 4.4 ensures that the latter is equivalent to that there exist $\mathbf{a} \in \partial\psi(\mathbf{B}'\mathcal{B}\mathbf{u}^*)$ and $\mathbf{b} \in \mathcal{N}(\mathbf{B}^\top)$ such that (24) and (25) hold. Note that $\mathbf{B}'\mathcal{B}\mathbf{u}^* = \mathbf{u}^*$, from which the desired result is obtained. When $\operatorname{rank}(\mathbf{B}) = m$, there holds $\mathcal{N}(\mathbf{B}^\top) = (\mathcal{R}(\mathbf{B}))^\perp = \{0\}$. It follows that $\mathbf{b}$ in (24) and (25) is the zero vector. Thus, (27) and (28) can be obtained directly. □

For the most sparse solution $\mathbf{u}^*$ under $\mathbf{B}$ (that is, $\mathbf{B}\mathbf{u}^* = 0$), conditions (24) and (25) reduce to $(\widetilde{\mathbf{B}}'_2)^\top\mathbf{a} = 0$ and $\lambda \geqslant \left\|(\widetilde{\mathbf{B}}'_1)^\top\mathbf{a} + \mathbf{b}\right\|_\infty$, where $\widetilde{\mathbf{B}}'_1$ and $\widetilde{\mathbf{B}}'_2$ denote the matrices composed of the first m columns and the last $n-r$ columns of $\mathbf{B}'$, respectively.

When ψ is differentiable, theorem 4.5 has the following simple form.

Corollary 4.6. *Suppose that ψ is a differentiable and convex function on $\mathbb{R}^n$. Then the regularization problem (1) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{m+1}$ if and only if there exists $\mathbf{b} \in \mathcal{N}(\mathbf{B}^\top)$ such that*

$$(\mathbf{B}'_j)^\top\nabla\psi(\mathbf{u}^*) = 0, j\in\mathbb{N}_{m+n-r}\setminus\mathbb{N}_m,\ \ \lambda = -((\mathbf{B}'_{k_i})^\top\nabla\psi(\mathbf{u}^*) + b_{k_i})\operatorname{sign}(z^*_{k_i}), i\in\mathbb{N}_l,$$

$$\lambda \geqslant \left|(\mathbf{B}'_j)^\top\nabla\psi(\mathbf{u}^*) + b_j\right|,\ j\in\mathbb{N}_m\setminus\{k_i : i\in\mathbb{N}_l\}.$$

In particular, if $\operatorname{rank}(\mathbf{B}) = m$, then the conditions reduce to

$$(\mathbf{B}'_j)^\top\nabla\psi(\mathbf{u}^*) = 0, j\in\mathbb{N}_n\setminus\mathbb{N}_m,\ \ \lambda = -(\mathbf{B}'_{k_i})^\top\nabla\psi(\mathbf{u}^*)\operatorname{sign}(z^*_{k_i}), i\in\mathbb{N}_l, \tag{29}$$

$$\lambda \geqslant \left|(\mathbf{B}'_j)^\top\nabla\psi(\mathbf{u}^*)\right|,\ j\in\mathbb{N}_m\setminus\{k_i : i\in\mathbb{N}_l\}. \tag{30}$$

In the remaining part of this section, we apply theorem 4.5 or corollary 4.6 to several specific models described in section 2. We first consider the ℓ_1 SVM classification model with the hinge loss function

$$\min\left\{\phi(\mathbf{Y}\mathbf{K}'\mathbf{u}) + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u}\in\mathbb{R}^{n+1}\right\}. \tag{31}$$

By introducing a univariate function $\phi(x) := \max\{1-x, 0\}$, $x\in\mathbb{R}$, we represent $\boldsymbol{\phi}$ as $\boldsymbol{\phi}(\mathbf{x}) = \sum_{j\in\mathbb{N}_n}\phi(x_j)$, for all $\mathbf{x} := [x_j : j\in\mathbb{N}_n] \in \mathbb{R}^n$.

Corollary 4.7. *The problem (31) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exists $\mathbf{c} := [c_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$ with $c_j \in \partial\phi((\mathbf{Y}\mathbf{K}'\mathbf{u}^*)_j)$, $j \in \mathbb{N}_n$, such that*

$$\mathbf{y}^\top\mathbf{c} = 0, \lambda = -(\mathbf{Y}\mathbf{K}_{k_i})^\top\mathbf{c}\,\mathrm{sign}(z^*_{k_i}), i \in \mathbb{N}_l,\ \lambda \geqslant \left|(\mathbf{Y}\mathbf{K}_j)^\top\mathbf{c}\right|, j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{32}$$

Proof. Clearly, the fidelity term $\psi(\mathbf{u}) := \phi(\mathbf{Y}\mathbf{K}'\mathbf{u})$, $\mathbf{u} \in \mathbb{R}^{n+1}$ is convex on $\mathbb{R}^{n+1}$ and the matrix $\mathbf{B} := [\mathbf{I}_n\ 0]$ has full row rank. By theorem 4.5, problem (31) has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exists $\mathbf{a} \in \partial\psi(\mathbf{u}^*)$ such that (27) and (28) with m, n being replaced by $n, n+1$, respectively, hold.

It remains to verify that in this case (27) and (28) reduce to (32). We first describe the subdifferential of ψ. By the chain rule of the subdifferential, we have for all $\mathbf{u} \in \mathbb{R}^{n+1}$ that $\partial\psi(\mathbf{u}) = (\mathbf{Y}\mathbf{K}')^\top\partial\phi(\mathbf{Y}\mathbf{K}'\mathbf{u})$. It follows from the separable representation of ϕ that $\partial\phi(\mathbf{x}) = \{\mathbf{c} := [c_j : j \in \mathbb{N}_n] \in \mathbb{R}^n : c_j \in \partial\phi(x_j), j \in \mathbb{N}_n\}$, for all $\mathbf{x} := [x_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$. Substituting the above equation with $\mathbf{x} := \mathbf{Y}\mathbf{K}'\mathbf{u}$ into the subdifferential of ψ, we obtain that $\partial\psi(\mathbf{u}) = \{(\mathbf{Y}\mathbf{K}')^\top\mathbf{c} : \mathbf{c} := [c_j : j \in \mathbb{N}_n] \in \mathbb{R}^n, c_j \in \partial\phi((\mathbf{Y}\mathbf{K}'\mathbf{u})_j),\ j \in \mathbb{N}_n\}$. We next represent the matrix $\mathbf{B}'$. Note that $\mathbf{B}$ has the SVD $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ with $\mathbf{U} := \mathbf{I}_n$, $\mathbf{\Lambda} := \mathbf{B}$ and $\mathbf{V} := \mathbf{I}_{n+1}$. It follows that $\mathbf{B}' = \mathbf{I}_{n+1}$. Substituting the representations of $\mathbf{B}'$ and $\partial\psi$ into (27) and (28) and noting that $(\mathbf{B}'_j)^\top(\mathbf{Y}\mathbf{K}')^\top = (\mathbf{Y}\mathbf{K}_j)^\top$ for all $j \in \mathbb{N}_n$ and $(\mathbf{B}'_{n+1})^\top(\mathbf{Y}\mathbf{K}')^\top = \mathbf{y}^\top$, we get the desired conditions (32). □

We next consider the ℓ_1 SVM regression model with the ϵ-insensitive loss function

$$\min\left\{\phi_{\mathbf{y},\epsilon}(\mathbf{K}'\mathbf{u}) + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^{n+1}\right\}, \tag{33}$$

The function $\phi_{\mathbf{y},\epsilon}$ is additively separable with the form $\phi_{\mathbf{y},\epsilon}(\mathbf{x}) = \sum_{j\in\mathbb{N}_n}\phi_{y_j,\epsilon}(x_j)$, $\mathbf{x} := [x_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$, where $\phi_{y,\epsilon}(t) := \max\{|y-t| - \epsilon, 0\}$, $t \in \mathbb{R}$. The following characterization may be proved by theorem 4.5 and arguments similar to those used in the proof of corollary 4.7. We omit the details of the proof.

Corollary 4.8. *The regularization problem (33) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there exists $\mathbf{c} := [c_j : j \in \mathbb{N}_n] \in \mathbb{R}^n$ with $c_j \in \partial\phi_{y_j,\epsilon}((\mathbf{K}'\mathbf{u}^*)_j)$, $j \in \mathbb{N}_n$, such that*

$$1_n^\top\mathbf{c} = 0,\ \lambda = -(\mathbf{K}_{k_i})^\top\mathbf{c}\mathrm{sign}(z^*_{k_i}),\ i \in \mathbb{N}_l,\ \ \lambda \geqslant \left|(\mathbf{K}_j)^\top\mathbf{c}\right|, j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{34}$$

The following example concerns the total-variation signal denoising model

$$\min\left\{\frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{D}^{(1)}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^n\right\}. \tag{35}$$

Suppose that $\mathbf{D}^{(1)}$ has the SVD $\mathbf{D}^{(1)} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ and $\mathbf{D}^{(1)'} := \mathbf{V}\mathbf{\Lambda}'\mathbf{U}'$. It follows from [53] that $\mathbf{V}_n = \frac{\sqrt{n}}{n}1_n$, which together with the definition of $\mathbf{D}^{(1)'}$ leads to $\mathbf{D}_n^{(1)'} = \frac{\sqrt{n}}{n}1_n$.

Corollary 4.9. *The regularization problem (35) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{D}^{(1)}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_n$ if and only if*

$$1_n^\top(\mathbf{u}^* - \mathbf{x}) = 0,\ \ \lambda = (\mathbf{D}_{k_i}^{(1)'})^\top(\mathbf{x} - \mathbf{u}^*)\mathrm{sign}(z^*_{k_i}),\ i \in \mathbb{N}_l,$$

$$\lambda \geqslant \left|(\mathbf{D}_j^{(1)'})^\top(\mathbf{u}^* - \mathbf{x})\right|, j \in \mathbb{N}_{n-1} \setminus \{k_i : i \in \mathbb{N}_l\}. \tag{36}$$

Proof. Since the fidelity term $\psi := \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2$ is differentiable and convex and the matrix $\mathbf{D}^{(1)}$ has full row rank, we may prove the results in this corollary by corollary 4.6. It is done by substituting $\nabla\psi(\mathbf{u}^*) = \mathbf{u}^* - \mathbf{x}$ and $\mathbf{D}_n^{(1)'} = \frac{\sqrt{n}}{n}1_n$ into (29) and (30). □

A parameter choice strategy for the most sparse solution under the transform $\mathbf{D}^{(1)}$ is provided in following remark, whose proof is included in appendix. We denote by $\widetilde{\mathbf{D}}^{(1)'}$ the matrix composed of the first $n-1$ columns of $\mathbf{D}^{(1)'}$.

Remark 4.10. The regularization problem (35) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{D}^{(1)}\mathbf{u}^* = 0$ if and only if $\lambda \geqslant \left\|(\widetilde{\mathbf{D}}^{(1)'})^\top \mathbf{x}\right\|_\infty$. Moreover, the solution $\mathbf{u}^*$ with $\mathbf{D}^{(1)}\mathbf{u}^* = 0$ has the form $\mathbf{u}^* := \frac{1}{n} 1_n^\top \mathbf{x} 1_n$.

We also consider the ℓ_1 SVM models for classification/regression with the squared loss function (6). As pointed out in section 2, these models can be formulated as

$$\min\left\{\frac{1}{2}\|\mathbf{K}'\mathbf{u} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^{n+1}\right\}. \tag{37}$$

By employing corollary 4.6 and arguments similar to those used in the proof of corollary 4.9, we obtain the characterization as follows.

Corollary 4.11. *The regularization problem (37) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{n+1}$ if and only if there hold*

$$1_n^\top(\mathbf{K}'\mathbf{u}^* - \mathbf{y}) = 0,\ \ \lambda = (\mathbf{K}_{k_i})^\top(\mathbf{y} - \mathbf{K}'\mathbf{u}^*)\,\mathrm{sign}(z^*_{k_i}),\ i \in \mathbb{N}_l,$$
$$\lambda \geqslant \left|(\mathbf{K}_j)^\top(\mathbf{K}'\mathbf{u}^* - \mathbf{y})\right|,\ j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}.$$

When the solution has the most sparsity under the transform $\mathbf{B}$, the characterization stated in corollary 4.11 reduces to a simple form.

Remark 4.12. The regularization problem (37) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* = 0$ if and only if $\lambda \geqslant \left\|\mathbf{K}^\top\left(\frac{1}{n}1_n^\top\mathbf{y}1_n - \mathbf{y}\right)\right\|_\infty$. Moreover, the solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* = 0$ has the form $\mathbf{u}^* := \begin{bmatrix} 0 \\ \frac{1}{n}1_n^\top\mathbf{y} \end{bmatrix}$.

We finally consider the ℓ_1 regularized logistic regression model

$$\min\left\{\phi(\mathbf{Y}\mathbf{X}'\mathbf{u}) + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^{d+1}\right\}. \tag{38}$$

The proof of the following result is similar to that of corollary 4.9 and thus is omitted. For each $\mathbf{u} \in \mathbb{R}^{d+1}$, we set $\mathbf{c}_\mathbf{u} := [(1 + \exp((\mathbf{Y}\mathbf{X}'\mathbf{u})_j))^{-1} : j \in \mathbb{N}_n] \in \mathbb{R}^n$.

Corollary 4.13. *The regularization problem (38) with $\lambda > 0$ has a solution $\mathbf{u}^*$ and $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i} \in \Omega_l$ for some $l \in \mathbb{Z}_{d+1}$ if and only if there hold $\mathbf{y}^\top\mathbf{c}_{\mathbf{u}^*} = 0$ and*

$$\lambda = \frac{1}{n}(\mathbf{Y}\mathbf{X}_{k_i})^\top\mathbf{c}_{\mathbf{u}^*}\mathrm{sign}(z^*_{k_i}),\ i \in \mathbb{N}_l,\ \lambda \geqslant \frac{1}{n}\left|(\mathbf{Y}\mathbf{X}_j)^\top\mathbf{c}_{\mathbf{u}^*}\right|,\ j \in \mathbb{N}_d \setminus \{k_i : i \in \mathbb{N}_l\}.$$

When the most sparse solution under the transform $\mathbf{B}$ is desired, we have the parameter choice strategy described in the next remark. We denote by n_+ and n_- the numbers of data with output $y_j = 1$ and $y_j = -1$, respectively, and set $\mathbf{c} := [(1 + (n_+/n_-)^{y_j})^{-1} : j \in \mathbb{N}_n] \in \mathbb{R}^n$.

Remark 4.14. The regularization problem (38) with $\lambda > 0$ has a solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* = 0$ if and only if $\lambda \geqslant \frac{1}{n}\left\|(\mathbf{Y}\mathbf{X})^\top\mathbf{c}\right\|_\infty$. Moreover, the solution $\mathbf{u}^*$ with $\mathbf{B}\mathbf{u}^* = 0$ has the form $\mathbf{u}^* := \begin{bmatrix} 0 \\ \ln(n_+/n_-) \end{bmatrix}$.

We comment that the characterizations about sparsity of a solution of (38) under the transform $\mathbf{B}$, stated in corollary 4.13 and remark 4.14, were established in [33].

We return to the general case. Theorem 4.5 and corollary 4.6 can not be used directly as a parameter choice strategy since the characterizations involve the unknown solution. In the

rest of this section, we develop parameter choice strategies from the characterizations. Suppose that ψ is differentiable and $\mathbf{B}$ has full row rank. We first consider the case when partial information of the solution is known. The indices $k_i, i \in \mathbb{N}_l$, which appear in equalities (29) for the parameter λ to satisfy, correspond to the non-zero components of the solution $\mathbf{u}^*$, while those which appear in inequalities (30) correspond to the zero components of the solution. By enlarging the lower bounds of the inequalities, we can obtain a sufficient condition for the solution to have sparsity of a certain level under the transform $\mathbf{B}$, which gives us a way to choose the parameter. Along this direction, we get the following result.

Proposition 4.15. *Let ψ be a differentiable and convex function on $\mathbb{R}^n$ and $\boldsymbol{B}$ is a real $m \times n$ matrix having full row rank. Suppose that for each $j \in \mathbb{N}_m$, there exists $L_j > 0$ such that $\left|(\mathbf{B}_j')^\top(\nabla\psi(\mathbf{u}) - \nabla\psi(\mathbf{w}))\right| \leqslant L_j\|\mathbf{u}-\mathbf{w}\|_2$, for all $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$. Let $\mathbf{u}^* \in \mathbb{R}^n$ be a solution of problem (1) with $\lambda > 0$ and $\mathbf{v} \in \mathbb{R}^n$ satisfy $\|\mathbf{u}^* - \mathbf{v}\|_2 \leqslant \epsilon$ for some $\epsilon > 0$. If there exist distinct $k_i \in \mathbb{N}_m$, $i \in \mathbb{N}_l$, for some $l \in \mathbb{Z}_{m+1}$ such that*

$$\lambda > \left|(\mathbf{B}_j')^\top\nabla\psi(\mathbf{v})\right| + \epsilon L_j, \text{ for all } j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}, \tag{39}$$

then the sparsity level of $\mathbf{u}^*$ under the transform $\boldsymbol{B}$ is less than or equal to l.

Proof. It follows from the assumption that $\left|(\mathbf{B}_j')^\top(\nabla\psi(\mathbf{u}^*) - \nabla\psi(\mathbf{v}))\right| \leqslant L_j\|\mathbf{u}^* - \mathbf{v}\|_2$, for all $j \in \mathbb{N}_m$, which together with $\|\mathbf{u}^* - \mathbf{v}\|_2 \leqslant \epsilon$ leads to $\left|(\mathbf{B}_j')^\top(\nabla\psi(\mathbf{u}^*) - \nabla\psi(\mathbf{v}))\right| \leqslant \epsilon L_j$, for all $j \in \mathbb{N}_m$. Substituting these inequalities into inequality (39), we get that $\lambda > \left|(\mathbf{B}_j')^\top\nabla\psi(\mathbf{v})\right| + \left|(\mathbf{B}_j')^\top(\nabla\psi(\mathbf{u}^*) - \nabla\psi(\mathbf{v}))\right|$, for all $j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}$. This further yields that $\lambda > \left|(\mathbf{B}_j')^\top\nabla\psi(\mathbf{u}^*)\right|$, for all $j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}$. Suppose that $\mathbf{B}\mathbf{u}^* := \sum_{j\in\mathbb{N}_m} z_j^*\mathbf{e}_j$. By corollary 4.6, from the above inequality we conclude that $z_j^* = 0$ for all $j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}$. In fact, if there exists $j_0 \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_l\}$ such that $z_{j_0}^* \neq 0$, then corollary 4.6 ensures that $\lambda = -(\mathbf{B}_{j_0}')^\top\nabla\psi(\mathbf{u}^*)\text{sign}(z_{j_0}^*)$. That is, $\lambda = \left|(\mathbf{B}_{j_0}')^\top\nabla\psi(\mathbf{u}^*)\right|$, which contradicts to $\lambda > \left|(\mathbf{B}_{j_0}')^\top\nabla\psi(\mathbf{u}^*)\right|$. Thus, we conclude that the sparsity level of $\mathbf{u}^*$ under the transform $\mathbf{B}$ is less than or equal to l. □

We now develop an iterative scheme based on the characterization described in corollary 4.6 for choosing a parameter λ with which a solution of problem (1) has sparsity of a prescribed level under the transform $\mathbf{B}$. We need the following result derived from corollary 4.6.

Proposition 4.16. *Suppose that $\mathbf{u}^*$ is a solution of problem (1) with $\lambda^* > 0$ having sparsity of level $l^* \in \mathbb{Z}_{m+1}$ under the transform $\boldsymbol{B}$. Let $a_j(\mathbf{u}^*) := \left|(\mathbf{B}_j')^\top\nabla\psi(\mathbf{u}^*)\right|$, $j \in \mathbb{N}_m$. If they are rearranged in a nondecreasing order: $a_{j_1}(\mathbf{u}^*) \leqslant \cdots \leqslant a_{j_m}(\mathbf{u}^*)$ with distinct $j_i \in \mathbb{N}_m$, $i \in \mathbb{N}_m$, then*

$$a_{j_1}(\mathbf{u}^*) \leqslant \cdots \leqslant a_{j_{m-l^*}}(\mathbf{u}^*) \leqslant \lambda^* = a_{j_{m-l^*+1}}(\mathbf{u}^*) = \cdots = a_{j_m}(\mathbf{u}^*). \tag{40}$$

Proof. Since $\mathbf{u}^*$ is a solution of problem (1) with $\lambda^* > 0$ having sparsity of level $l^* \in \mathbb{Z}_{m+1}$ under the transform $\mathbf{B}$, there are l^* distinct integers $k_i \in \mathbb{N}_m$, $i \in \mathbb{N}_{l^*}$, such that $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_{l^*}} z_{k_i}^*\mathbf{e}_{k_i} \in \Omega_{l^*}$. By corollary 4.6, we find that (29) and (30) hold. Because $\lambda^* > 0$, the second equality of (29) implies that $\lambda^* = a_{k_i}(\mathbf{u}^*)$, for all $i \in \mathbb{N}_{l^*}$. Moreover, (30) gives $\lambda^* \geqslant a_j(\mathbf{u}^*)$, for all $j \in \mathbb{N}_m \setminus \{k_i : i \in \mathbb{N}_{l^*}\}$. By rearranging the sequence $a_j(\mathbf{u}^*)$, $j \in \mathbb{N}_m$, in a nondecreasing order, we get the desired result. □

We next propose an iterative scheme which simultaneously determines parameter λ^* and solution $\mathbf{u}^*$ having a prescribed sparsity level l^* under the transform $\mathbf{B}$. The iteration begins with an initial λ that is large enough to ensure the sparsity level l under the transform $\mathbf{B}$ of the solution $\mathbf{u}$ of model (1) with this λ does not exceed l^*. Proposition 4.16 which shows that the desired λ^* and its corresponding $\mathbf{u}^*$ must satisfy (40) motivates us to update λ by taking

Algorithm 1. Parameter choice for the regularization problem (1).

Input: ψ, $\mathbf{B}$, l^*.
Initialize: choose an initial λ large enough that guarantees $l \leqslant l^*$.
Repeat :
- Solve model (1) with λ for $\mathbf{u}$ and count the sparsity level l of $\mathbf{Bu}$.
- If $l > l^*$, initialize $s := 0$.
  Repeat :
  - Update $s := s + l - l^*$ and $\lambda := \min\left\{a_{j_{m-l^*+s}}, a\right\}$.
  - Solve model (1) with λ for $\mathbf{u}$ and count the sparsity level l of $\mathbf{Bu}$.

  Until $l \leqslant l^*$.
- If $l < l^*$, do the following steps:
  - Compute $a_j := \left|(\mathbf{B}_j')^\top \nabla\psi(\mathbf{u})\right|, j \in \mathbb{N}_m$.
  - Sort: $a_{j_1} \leqslant \cdots \leqslant a_{j_m}$ with distinct $j_i \in \mathbb{N}_m$, $i \in \mathbb{N}_m$.
  - Compute $a := \max\left\{a_j : a_j < \lambda, j \in \mathbb{N}_m\right\}$.
  - Update $\lambda := \min\left\{a_{j_{m-l^*}}, a\right\}$.

Until $l = l^*$.
Output: $\lambda^* := \lambda$, $\mathbf{u}^* := \mathbf{u}$.

the $(m-l^*)$th element among the ordered sequence $a_{j_i}(\mathbf{u})$, $i \in \mathbb{N}_m$. Suppose that at step k, we have a λ^k and the corresponding solution $\mathbf{u}^k$ with the sparsity level l^k under the transform $\mathbf{B}$, satisfying

$$a_{j_1}(\mathbf{u}^k) \leqslant \cdots \leqslant a_{j_{m-l^k}}(\mathbf{u}^k) \leqslant \lambda^k = a_{j_{m-l^k+1}}(\mathbf{u}^k) = \cdots = a_{j_m}(\mathbf{u}^k), \tag{41}$$

according to proposition 4.16. If $l^k = l^*$, the iteration terminates with the desired parameter λ^* and solution $\mathbf{u}^*$ with sparsity level l^* under the transform $\mathbf{B}$. Otherwise, we continue the iteration. If $l^k > l^*$, this indicates that λ^k is too small and thus in our next step of iteration we should choose λ^{k+1} greater than λ^k. However, by (41) all elements in the sequence $a_{j_k}(\mathbf{u}^k)$, $k \in \mathbb{N}_m$, are less than or equal to λ^k. Thus, we cannot choose a desired parameter λ from the sequence and we should go back to the sequence in step $k-1$ to find an appropriate parameter. If $l^k < l^*$, this indicates that λ^k is too large and thus in our next step of iteration we should choose λ^{k+1} from one of $a_{j_1}(\mathbf{u}^k), \ldots, a_{j_{m-l^k}}(\mathbf{u}^k)$.

We summarize the iterative scheme in algorithm 1. Numerical examples presented in section 6 show that algorithm 1 converges and converges fast. Theoretical analysis of algorithm 1 is postponed to a future occasion.

5. Parameter choices for alleviating the ill-posedness and promoting sparsity of the regularized solutions

As we pointed out earlier, the purpose of imposing the ℓ_1 regularization is two-folds: alleviating the ill-posedness when given data are noisy and promoting sparsity of a regularized solution. In this section, we demonstrate how the regularization parameter λ can be chosen to achieve both of these by considering a lasso regularized model.

We aim at a prediction $\mathbf{u} \in \mathbb{R}^n$ from a given response vector $\mathbf{x} \in \mathbb{R}^p$, via the equation $\mathbf{Au} = \mathbf{x}$. We recover $\mathbf{u}$ from a noisy response $\mathbf{x}^\delta$ (instead of $\mathbf{x}$) with $\|\mathbf{x}^\delta - \mathbf{x}\|_2 \leqslant \delta$ for a given noise level δ by the lasso regularized model

$$\min\left\{\frac{1}{2}\|\mathbf{Au} - \mathbf{x}^\delta\|_2^2 + \lambda\|\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^n\right\}. \tag{42}$$

Appropriate choice of the matrix $\mathbf{A}$ can enrich the representation of the solution $\mathbf{u}_\lambda^\delta$ and thus help avoid underfitting. The regularization term is added not only to overcome overfitting [2, 9, 21, 45, 47] when data contaminates noise but also to impose sparsity of $\mathbf{u}_\lambda^\delta$. We are interested in choices of the parameter λ that balances the error of $\mathbf{u}_\lambda^\delta$ and its sparsity. Here, the error is compared to the minimal norm solution

$$\tilde{\mathbf{u}} := \operatorname{argmin}\{\|\mathbf{u}\|_1 : \mathbf{A}\mathbf{u} = \mathbf{x}, \mathbf{u} \in \mathbb{R}^n\} \tag{43}$$

of the prediction problem. The minimal norm problem itself is a recent research topic of great interest [12, 17, 25]. We assume that (43) has a unique solution $\tilde{\mathbf{u}} := [\tilde{u}_j : j \in \mathbb{N}_n]$. We briefly review the uniqueness of $\tilde{\mathbf{u}}$ presented in [25]. Denote by J the support of $\tilde{\mathbf{u}}$ and let $J^c := \mathbb{N}_n \setminus J$. We let $\mathbf{v} := [\operatorname{sign}(\tilde{u}_j) : j \in J]$, $\mathbf{A}' := [\mathbf{A}_j : j \in J]$ and $\mathbf{A}'' := [\mathbf{A}_j : j \in J^c]$. It is known from [25] that (43) has a unique solution $\tilde{\mathbf{u}}$ if and only if $\mathbf{A}\tilde{\mathbf{u}} = \mathbf{x}$, $\mathbf{A}'$ has full column rank and there exists $\mathbf{y} \in \mathbb{R}^p$ such that $(\mathbf{A}')^\top \mathbf{y} = \mathbf{v}$ and $\|(\mathbf{A}'')^\top \mathbf{y}\|_\infty < 1$.

Below, we state an error estimate between $\mathbf{u}_\lambda^\delta$ and $\tilde{\mathbf{u}}$ obtained by specializing a general argument (proposition 8 and theorem 15 of [26]) to problem (42). The uniqueness of the solution $\tilde{\mathbf{u}}$ ensures that there exists $\mathbf{y} \in \mathbb{R}^p$ such that $\|(\mathbf{A}'')^\top \mathbf{y}\|_\infty < 1$. Using the vector $\mathbf{y}$, constants β_1 and β_2 independent of δ and λ were introduced in [26]

$$\beta_1 := \frac{1 - \|(\mathbf{A}'')^\top \mathbf{y}\|_\infty}{1 + \|(\mathbf{A}')^\dagger\|_2 \|\mathbf{A}\|_2}, \quad \beta_2 := \frac{\|(\mathbf{A}')^\dagger\|_2 (1 - \|(\mathbf{A}'')^\top \mathbf{y}\|_\infty)}{1 + \|(\mathbf{A}')^\dagger\|_2 \|\mathbf{A}\|_2} + \|\mathbf{y}\|_2.$$

Throughout this section, we assume that $\mathbf{A} \in \mathbb{R}^{p\times n}$, the minimal norm problem (43) with $\mathbf{x} \in \mathbb{R}^p$ has a unique solution $\tilde{\mathbf{u}}$, and for $\delta > 0$, $\mathbf{x}^\delta \in \mathbb{R}^p$ satisfies $\|\mathbf{x}^\delta - \mathbf{x}\|_2 \leqslant \delta$.

Lemma 5.1. *If $\mathbf{u}_\lambda^\delta$ is a solution of the regularization problem (42), then for all $\delta, \lambda > 0$*

$$\left\|\mathbf{u}_\lambda^\delta - \tilde{\mathbf{u}}\right\|_2 \leqslant \frac{\lambda\beta_2^2}{\beta_1} + \frac{\delta^2}{2\lambda\beta_1} + \frac{\beta_2\delta}{\beta_1}. \tag{44}$$

We first consider the case that the predictor matrix $\mathbf{A} \in \mathbb{R}^{p\times n}$ is $\mathcal{S}$-block separable, that is, it satisfies condition (14) with respect to the partition $\mathcal{S} := \{S_1, S_2, \ldots, S_d\}$ of the set $\mathbb{N}_n$. Let $a_j^\delta := \left\|(\mathbf{A}_{(j)})^\top \mathbf{x}^\delta\right\|_\infty$, $j \in \mathbb{N}_d$, rearranged in a nondecreasing order: $a_{k_1}^\delta \leqslant a_{k_2}^\delta \leqslant \cdots \leqslant a_{k_d}^\delta$ with distinct $k_i \in \mathbb{N}_d$, $i \in \mathbb{N}_d$. We now present our results.

Theorem 5.2. *(a) If $\lambda := a_{k_{d-l}}^\delta$ for a given $l \in \mathbb{Z}_{d+1}$, then problem (42) has a sparse solution $\mathbf{u}_\lambda^\delta$ with the $\mathcal{S}$-block sparsity of level $\leqslant l$ satisfying the error bound*

$$\left\|\mathbf{u}_\lambda^\delta - \tilde{\mathbf{u}}\right\|_2 \leqslant \frac{a_{k_{d-l}}^\delta \beta_2^2}{\beta_1} + \frac{\delta^2}{2a_{k_{d-l}}^\delta \beta_1} + \frac{\beta_2\delta}{\beta_1}. \tag{45}$$

(b) If $\lambda := C\delta$ for a constant $C > 0$ such that $a_{k_{d-l}}^\delta \leqslant C\delta < a_{k_{d-l+1}}^\delta$ for some $l \in \mathbb{Z}_{d+1}$, then (42) has a solution $\mathbf{u}_\lambda^\delta$ with the $\mathcal{S}$-block sparsity of level l satisfying

$$\left\|\mathbf{u}_\lambda^\delta - \tilde{\mathbf{u}}\right\|_2 \leqslant C'\delta, \quad \text{where} \quad C' := (2C^2\beta_2^2 + 2C\beta_2 + 1)/(2C\beta_1). \tag{46}$$

Proof. Since $\mathbf{A} \in \mathbb{R}^{p\times n}$ is $\mathcal{S}$-block separable, condition (14) is satisfied. We then prove this theorem by employing corollary 3.5 with $\mathbf{x}$ being replaced by $\mathbf{x}^\delta$. Moreover, since (43) has a unique solution $\tilde{\mathbf{u}}$, lemma 5.1 ensures that the error estimate (44) holds.

We first prove statement (a). Since the parameter is chosen as $\lambda := a_{k_{d-l}}^\delta$, according to the order $a_{k_1}^\delta \leqslant \cdots \leqslant a_{k_d}^\delta$ of the sequence $a_{k_j}, j \in \mathbb{N}_d$, we have that $\lambda \geqslant a_{k_j}^\delta$, for all $j \in \mathbb{N}_{d-l}$. Appealing to corollary 3.5, we conclude that (42) with λ so chosen has a solution $\mathbf{u}_\lambda^\delta$ with the $\mathcal{S}$-block sparsity of level $\leqslant l$. The error bound (45) of $\mathbf{u}_\lambda^\delta$ is obtained by substituting $\lambda = a_{k_{d-l}}^\delta$ into the right hand side of estimate (44).

We next show statement (b). Substituting $\lambda = C\delta$ into the right hand side of the estimate (44) with straightforward computation leads to the error bound (46). In addition, since λ is chosen so that $a^{\delta}_{k_{d-l}} \leqslant \lambda < a^{\delta}_{k_{d-l+1}}$ for an integer l, corollary 3.5 ensures that the corresponding solution $\mathbf{u}^{\delta}_{\lambda}$ of (42) has the $\mathcal{S}$-block sparsity of level l. □

Theorem 5.2, which extends the classical posterior parameter choice strategies [5, 26, 41, 46, 52, 65] for noisy data, provides parameter choice strategies which balance sparsity of the corresponding regularized solutions and their error bounds. Item (b) of theorem 5.2 shows that the proposed parameter choice strategy generates a regularized solution to have the same error bound as in [26] (overcoming overfitting that may be caused by noisy data) and sparsity of a prescribed level. We can obtain a special result when $\mathbf{A}$ is an orthogonal matrix of order n, where condition (14) holds for the nature partition $\mathcal{S} := \{S_1, \ldots, S_n\}$ of $\mathbb{N}_n$. It follows from the invertability of $\mathbf{A}$ that (43) has a unique solution. Hence, the parameter choice strategies follows directly from theorem 5.2 with the nature partition $\mathcal{S}$ of $\mathbb{N}_n$ and the sequence $a^{\delta}_j := \left\|(\mathbf{A}_j)^{\top}\mathbf{x}^{\delta}\right\|_{\infty}, j \in \mathbb{N}_n$.

We next consider the case when $\mathbf{A}$ can not be partitioned to satisfy condition (14). In this case, algorithm 1 provides a choice of the parameter with which problem (42) has a solution with a prescribed sparsity level. This together with corollary 3.8 and lemma 5.1 enables us to obtain parameter choice strategies balancing sparsity of the regularized solutions and their error bounds. For the solution $\mathbf{u}^{\delta}_{\lambda}$ of problem (42), we define the sequence $a^{\delta}_j(\mathbf{u}^{\delta}_{\lambda}) := \left|(\mathbf{A}_j)^{\top}(\mathbf{A}\mathbf{u}^{\delta}_{\lambda} - \mathbf{x}^{\delta})\right|, j \in \mathbb{N}_n$, rearranged in a nondecreasing order: $a^{\delta}_{k_1}(\mathbf{u}^{\delta}_{\lambda}) \leqslant \cdots \leqslant a^{\delta}_{k_n}(\mathbf{u}^{\delta}_{\lambda})$ with distinct $k_i \in \mathbb{N}_n, i \in \mathbb{N}_n$. The proof of the following theorem is similar to that of theorem 5.2 and it is left to the interested readers.

Theorem 5.3. *(a) If λ^* is chosen by algorithm 1 for a given $l^* \in \mathbb{Z}_{n+1}$, then the solution $\mathbf{u}^{\delta}_{\lambda^*}$ of problem (42) has sparsity of level l^* and satisfies the error bound*

$$\left\|\mathbf{u}^{\delta}_{\lambda^*} - \tilde{\mathbf{u}}\right\|_2 \leqslant \frac{\lambda^*\beta_2^2}{\beta_1} + \frac{\delta^2}{2\lambda^*\beta_1} + \frac{\beta_2\delta}{\beta_1}.$$

(b) If $\lambda := C\delta$ for a constant $C > 0$, then the solution $\mathbf{u}^{\delta}_{\lambda}$ of problem (42) has sparsity of level $l' \leqslant l$ where $l \in \mathbb{Z}_{n+1}$ satisfies $a^{\delta}_{k_{n-l}}(\mathbf{u}^{\delta}_{\lambda}) < C\delta \leqslant a^{\delta}_{k_{n-l+1}}(\mathbf{u}^{\delta}_{\lambda})$, and satisfies the error bound (46).

6. Numerical experiments

In this section, we present numerical results to validate the theory established in this paper. We have three types of numerical examples: (a) In sections 6.1–6.4, we validate the parameter choice strategy for a targeted sparsity level. (b) Examples in sections 6.5 and 6.6 confirm that the parameter choices can balance sparsity of the regularized solution and its approximation accuracy. (c) Those in sections 6.6 and 6.7 are designed to verify characterizations of the solutions of the regularization problems presented in theorem 4.5 and proposition 4.15. All the experiments are performed with Matlab R2018a on an Intel Core I9 (8-core) with 5.0 GHz and 32 Gb RAM.

In our numerical computation, the regularization problems are solved by the fixed point proximity algorithm (FPPA) developed in [3, 35, 43], which we review below. Suppose that $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is a convex function, with $\mathrm{dom}(f) := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) < +\infty\} \neq \emptyset$. The proximity operator $\mathrm{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ of f is defined by $\mathrm{prox}_f(\mathbf{x}) := \mathrm{argmin}\left\{\frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + f(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^n\right\}$, for $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\boldsymbol{\varphi} : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\boldsymbol{\omega} : \mathbb{R}^m \to \overline{\mathbb{R}}$ are two convex functions which may not be differentiable, and $\mathbf{C} \in \mathbb{R}^{m\times n}$. The optimization

Table 1. Parameter choices λ^* for targeted sparsity levels l^* for image denoising (total number of wavelet coefficients $n^2 := 65\,536$).

l^*	40 000	30 000	20 000	10 000	5000	500	0
λ^*	11.2188	16.5289	23.2754	34.3766	46.7302	205.5417	3707.6947
SL	40 000	30 000	20 000	10 000	5000	500	0
PSNR	25.5679	26.4841	26.8580	26.1921	25.0097	19.2129	5.5824

problem $\min\{\varphi(\mathbf{u}) + \omega(\mathbf{C}\mathbf{u}) : \mathbf{u} \in \mathbb{R}^n\}$ can be solved by FPPA: For given positive constants β, ρ and initial points $\mathbf{u}^0$, $\mathbf{z}^0$,

$$\begin{cases} \mathbf{u}^{k+1} = \operatorname{prox}_{\beta\varphi}\left(\mathbf{u}^k - \beta\mathbf{C}^\top\mathbf{z}^k\right), \\ \mathbf{z}^{k+1} = \rho\left(\mathcal{I} - \operatorname{prox}_{\frac{1}{\rho}\omega}\right)\left(\frac{1}{\rho}\mathbf{z}^k + \mathbf{C}\left(2\mathbf{u}^{k+1} - \mathbf{u}^k\right)\right). \end{cases} \tag{47}$$

According to [35], iteration (47) converges if $\beta\rho < 1/\|\mathbf{C}\|_2^2$. When the function φ is smooth, one can use the fast iterative shrinkage-thresholding algorithm (FISTA) [6] to speedup the convergence. In the numerical examples to be presented below, we obtain the solution $\mathbf{u}^*$ after iteration (47) converges. For convenience, we use 'SL' and 'BSL' to denote the sparsity level and the block sparsity level, respectively, of $\mathbf{u}^*$ (or $\mathbf{B}\mathbf{u}^*$).

6.1. Image denoising by a wavelet transform

Given a noisy image $\mathbf{x} \in \mathbb{R}^{n^2}$ and an orthogonal wavelet transform $\mathbf{W} \in \mathbb{R}^{n\times n}$, we consider in this experiment the image denoising model [22]

$$\min\left\{\frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{B}\mathbf{u}\|_1 : \mathbf{u} \in \mathbb{R}^{n^2}\right\}, \tag{48}$$

where $\mathbf{B} := \mathbf{W} \otimes \mathbf{W}$, with $\otimes$ denoting the Kronecker product. Note that the matrix $\mathbf{B}$, as a Kronecker product of two orthogonal matrices, is also orthogonal. By a change of variables $\mathbf{v} = \mathbf{B}\mathbf{u}$, we identify (48) as the lasso regularized model with $p = n$, and n, $\mathbf{A}$, $\mathbf{u}$ being replaced by n^2, $\mathbf{B}^\top$, $\mathbf{v}$, respectively, and with a separable fidelity term.

The experiment is conducted on gray scale test image 'Cameraman' with size 256×256. We use $\mathbf{f} := [f_j : j \in \mathbb{N}_{n^2}]$ with $n := 256$ for the original image, $\mathbf{x} := \mathbf{f} + \boldsymbol{\eta}$ for a noisy image, with noise $\boldsymbol{\eta}$ iid $N(0, \sigma^2)$ being the Gaussian noise at level $\sigma = 20$ and $\mathbf{W}$ being the Daubechies wavelet transform with the vanishing moments $\mathrm{N} = 4$ and the coarsest resolution level $\mathrm{L} = 4$. We set seven desired sparsity levels l^* from $\mathbb{N}_{n^2}$, which are $l^* := 40\,000, 30\,000, 20\,000, 10\,000, 5000, 500, 0$, and apply the parameter choice strategy described in corollary 3.5 to select regularization parameters λ^* such that the corresponding solutions enjoy the targeted sparsity levels l^*. Specifically, we rearrange the set $\{b_j : j \in \mathbb{N}_{n^2}\}$ with $b_j := \left|((\mathbf{B}^\top)_j)^\top\mathbf{x}\right|$, in a nondecreasing order: $b_{k_1} \leqslant b_{k_2} \leqslant \cdots \leqslant b_{k_{n^2}}$ with distinct $k_i \in \mathbb{N}_{n^2}, i \in \mathbb{N}_{n^2}$. For each of the targeted sparsity levels l^*, we choose the parameter as $\lambda^* := b_{k_{n^2-l^*}}$ and obtain a solution $\mathbf{v}^*$ by solving model (48) with $\mathbf{u}$ being replaced by $\mathbf{B}^\top\mathbf{v}$ and get the sparsity level SL of $\mathbf{v}^*$.

We report in table 1 the targeted sparsity levels l^*, the selected values of λ^*, the actual sparsity levels SL of the solutions $\mathbf{v}^*$ and the PSNR values of the denoised images $\mathbf{B}^\top\mathbf{v}^*$, where $\mathrm{PSNR} := 20\log_{10}\left(255 \times 256/\|\mathbf{f} - \mathbf{B}^\top\mathbf{v}^*\|_2\right)$, and show in figure 1 the original image, the noisy image, and the denoised image with $\lambda^* = 23.2754$. Numerical results in table 1 show that the sparsity levels SL of $\mathbf{v}^*$ coincide with the targeted sparsity levels l^*.

(a) (b) (c)

Figure 1. The 'Cameraman' image denoising: (a) the original image; (b) the noisy image with Gaussian noise at level $\sigma = 20$; (c) the denoised image with $\lambda = 23.2754$ (SL = 20 000, PSNR = 26.8580).

Table 2. Parameter choices λ^* for targeted $\mathcal{S}$-block sparsity levels l^* for signal denoising (total groups $d := 10$).

l^*	9	7	6	4	3	1	0
λ^*	0.1312	0.1400	0.1541	0.3667	0.9135	3.1193	5.0126
BSL	9	7	6	4	3	1	0
MSE	0.0017	0.0018	0.0020	0.0056	0.0161	0.0563	0.0858

6.2. Signal denoising by the group lasso regularized model

In this experiment, we consider the group lasso regularized model (4) with $\mathbf{A} \in \mathbb{R}^{n\times n}$ whose columns form an orthogonal wavelet basis. We recover the Doppler signal function

$$f(t) := \sqrt{t(1-t)}\sin\left((2.1\pi)/(t+0.05)\right),\ t \in [0,1], \tag{49}$$

from its noisy data by employing model (4). Let $n := 4096$. We generate sample points $t_j, j \in \mathbb{N}_n$, on a uniform grid in $[0,1]$ with step size $h := 1/(n-1)$ and consider recovering the signal $\mathbf{f} := [f(t_j) : j \in \mathbb{N}_n]$ from a noisy signal $\mathbf{x} := \mathbf{f} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an additive white Gaussian noise with the signal-to-noise ratio SNR = 7. The matrix $\mathbf{A}$ is the Daubechies wavelet transform with N := 6 and L := 3. We choose a partition $\mathcal{S} := \{S_1, S_2, \ldots, S_{10}\}$ of $\mathbb{N}_n$ with the cardinality $n_1 = 2^3$ and $n_j = 2^{j+1}, j \in \mathbb{N}_{10} \setminus \{1\}$, and decompose $\mathbf{A}$ into *ten* sub-matrices $\mathbf{A}_{(j)} := [\mathbf{A}_k : k \in S_j] \in \mathbb{R}^{n\times n_j}, j \in \mathbb{N}_{10}$. We set seven targeted $\mathcal{S}$-block sparsity levels $l^* := 9, 7, 6, 4, 3, 1, 0$. According to the parameter choice strategy stated in theorem 3.6, we select the parameter values λ^* with which model (4) has solutions having the targeted $\mathcal{S}$-block sparsity levels l^*. By rearranging the set $\{a_j : j \in \mathbb{N}_{10}\}$ with $a_j := \left\|(\mathbf{A}_{(j)})^\top \mathbf{x}\right\|_2 / \sqrt{n_j}$, in a nondecreasing order: $a_{k_1} \leqslant a_{k_2} \leqslant \cdots \leqslant a_{k_{10}}$ with distinct $k_i \in \mathbb{N}_{10}, i \in \mathbb{N}_{10}$, we choose $\lambda^* := a_{k_{10-l^*}}$. We solve model (4) with each selected value of λ^* for the corresponding solution $\mathbf{u}^*$.

The targeted $\mathcal{S}$-block sparsity levels l^*, the selected values of parameter λ^*, the actual $\mathcal{S}$-block sparsity levels BSL of $\mathbf{u}^*$ and the MSE values of the denoised signals $\mathbf{A}\mathbf{u}^*$ are reported in table 2, where MSE := $\frac{1}{n}\|\mathbf{f} - \mathbf{A}\mathbf{u}^*\|_2^2$. Observing from the numerical results, the actual BSL values of the solution $\mathbf{u}^*$ match exactly with the targeted $\mathcal{S}$-block sparsity levels l^*. Moreover, the approximation errors of the solutions corresponding to the selected values of λ^* exhibit increase as the values of λ^* become larger.

Table 3. Parameter choices λ^* for targeted sparsity levels l^* for signal denoising (total sparsity level $n := 4095$).

l^*	1488	744	355	160	67	12	0
λ^0	0.20	0.28	0.55	6.13	30.40	107.50	270.18
λ^*	0.1	0.2	0.5004	6.1277	30.2662	107.3430	270.18
SL	1488	744	355	160	67	12	0
NUM	8	11	33	34	75	78	1
MSE	0.0042	0.0020	0.0012	0.0078	0.0298	0.0657	0.0835

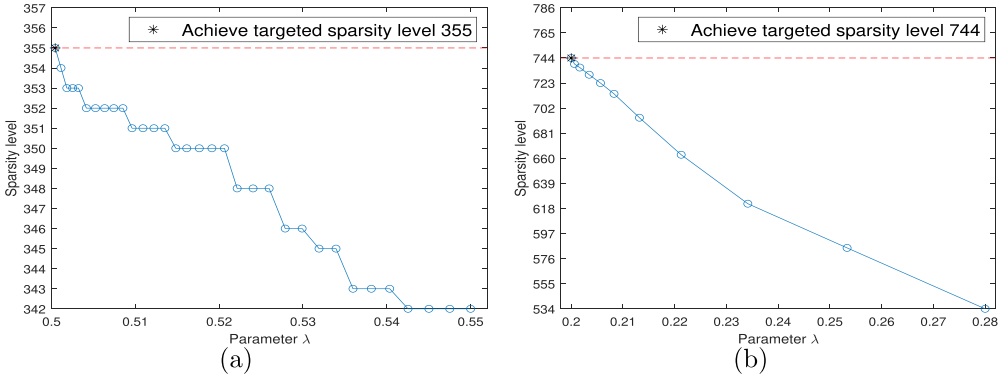

Figure 2. The parameter choice for targeted sparsity levels: (a) $l^* = 355$; (b) $l^* = 744$.

6.3. *Total-variation signal denoising*

We consider the total-variation signal denoising model (35), which is neither separable nor block separable. Again, we recover the Doppler signal function defined by (49) from its noisy data. The original signal $\mathbf{f}$ and the noisy signal $\mathbf{x}$ are chosen in the same way as in subsection 6.2. We set seven targeted sparsity levels l^* under the transform $\mathbf{D}^{(1)}$, that is, $l^* = 1488$, 744, 355, 160, 67, 12, 0. We apply algorithm 1 to find values of the parameter λ^* and the corresponding solutions $\mathbf{u}^*$ for the targeted sparsity levels l^* under the transform $\mathbf{D}^{(1)}$.

We report in table 3 the targeted sparsity levels l^*, initial values of λ^0, the selected values of parameter λ^*, the actual sparsity levels SL of $\mathbf{u}^*$ under the transform $\mathbf{D}^{(1)}$, the numbers NUM of updates for λ^* and the MSE values of the denoised signals $\mathbf{u}^*$, where $\text{MSE} := \frac{1}{n}\|\mathbf{f} - \mathbf{u}^*\|_2^2$. Moreover, we show in figure 2 the convergence process of algorithm 1 for the two prescribed sparsity levels $l^* = 355$ and $l^* = 744$. It demonstrates that algorithm 1 is convergent and converges fast. Thus, the proposed algorithm is not only effective but also computationally efficient.

6.4. *ℓ_1 SVM classification and regression with the squared loss function*

In this subsection, we present numerical results for solving the ℓ_1 SVM model (37) with the squared loss function for both classification and regression.

First, we consider the ℓ_1 SVM model (37) for classification. The dataset that we use for this experiment is the handwriting digits from the modified national institute of standards and technology (MNIST) database [34], which is composed of 60 000 training samples and 10 000 testing samples of the digits '0' through '9'. We study the binary classification problem with

Table 4. Parameter choices λ^* for targeted sparsity levels l^* for ℓ_1 SVM classification (total number of terms $n := 8141$).

l^*	6000	4000	2000	1000	500	100
λ^0	5	5	5	5	5	15
λ^*	0.0399	0.0916	0.3519	0.8959	1.7495	11.4520
SL	6005	4003	2005	1004	505	96
NUM	13	16	26	29	23	31
TrA	99.88%	99.84%	99.57%	99.07%	98.34%	96.31%
TeA	99.17%	99.12%	98.92%	98.72%	98.38%	96.81%

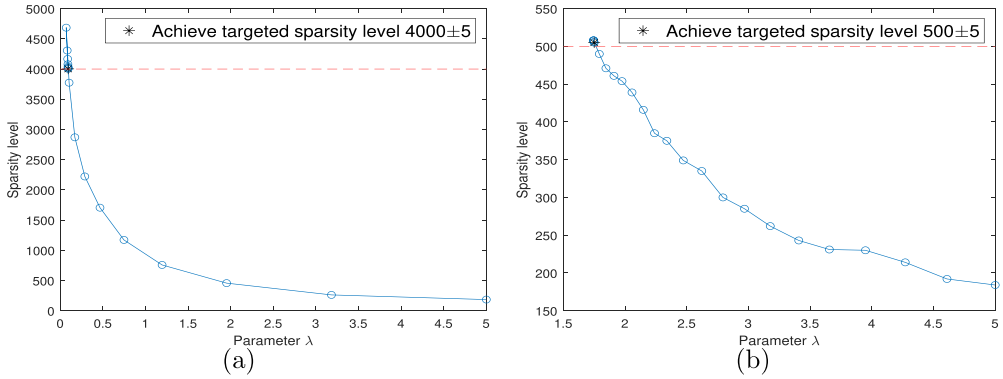

Figure 3. The parameter choice for targeted sparsity level l^*: (a) $l^* = 4000$; (b) $l^* = 500$.

two digits '7' and '9', by taking 8141 training samples and 2037 testing samples of these two digits from the database. The reason for which we choose these two particular digits is that it has been recognized that their handwriting is not easy to distinguish. The kernel we use for model (37) is the Gaussian kernel defined by

$$K(x,y) := \exp\left(-\|x-y\|_2^2/(2\mu^2)\right), \quad x,y \in \mathbb{R}^d, \tag{50}$$

with $\mu := 4.8$ and $d := 784$. Let $n := 8141$ be the number of training samples and $\mathbf{y} \in \{-1,1\}^n$ be the given vector storing labels of training data in which -1 and 1 represent the digits '7' and '9' respectively. We set six targeted sparsity levels l^* under the transform $\mathbf{B}$, that is, $l^* = 6000, 4000, 2000, 1000, 500, 100$, and apply algorithm 1 to find the value of parameter λ^* and the corresponding solution $\mathbf{u}^*$ for each l^*. In this experiment, we relax the stopping criteria for algorithm 1 to terminate if $|l - l^*| \leqslant 5$ instead of $l = l^*$.

We report in table 4 the targeted sparsity levels l^*, the initial values of λ^0, the selected values of parameter λ^*, the actual sparsity levels SL of the solutions $\mathbf{u}^*$ under the transform $\mathbf{B}$, the numbers NUM of updates for λ^*, the accuracy on the training datasets (TrA) and the accuracy on the testing datasets (TeA). Here, the accuracy is measured by labels that are correctly predicted by the model. Moreover, we show in figure 3 the convergence process of algorithm 1 for the two prescribed sparsity levels $l^* = 4000$ and $l^* = 500$. It shows that algorithm 1 is convergent and converges fast.

Secondly, we consider the ℓ_1 SVM model (37) for regression. The benchmark dataset is 'Mg' [13] with 1385 instances and each instance has six features. We take 1000 instances as training samples and 385 instances as testing samples. The kernel involved in model (37) is chosen as the Gaussian kernel defined by (50) with $\mu := 1.07$ and $d := 6$. Let $n := 1000$ be

Table 5. Parameter choices λ^* for targeted sparsity levels l^* for ℓ_1 SVM regression (total number of terms $n := 1000$).

l^*	600	500	400	300	200	100	50
λ^0	0.03	0.03	0.03	0.03	0.03	0.03	0.05
λ^*	0.0015	0.0020	0.0030	0.0052	0.0087	0.0238	0.0471
SL	601	502	402	301	197	100	47
NUM	12	10	10	9	12	6	3
TrMSE	0.0128	0.0129	0.0130	0.0132	0.0135	0.0140	0.0143
TeMSE	0.0146	0.0146	0.0147	0.0149	0.0150	0.0151	0.0151

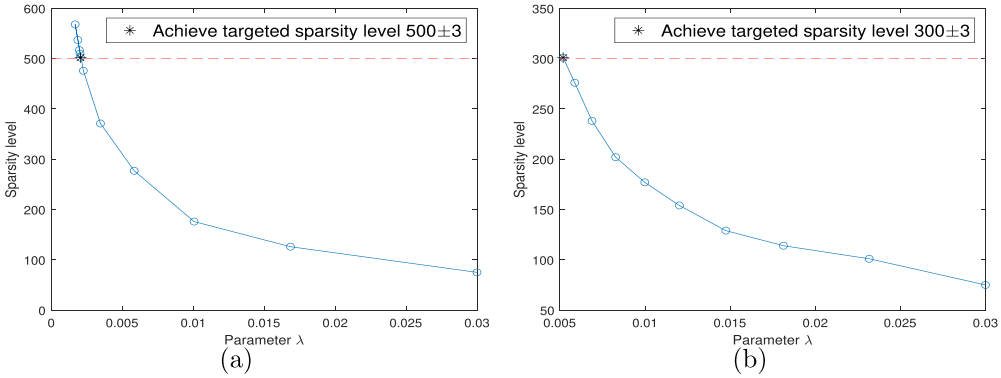

Figure 4. The parameter choice for targeted sparsity levels: (a) $l^* = 500$; (b) $l^* = 300$.

the number of training samples and $\mathbf{y} \in \mathbb{R}^n$ be the given labels. We set seven targeted sparsity levels l^* under the transform $\mathbf{B}$, which are $l^* = 600, 500, 400, 300, 200, 100, 50$. For each l^*, we use algorithm 1 to get the parameter λ^* and the corresponding solution $\mathbf{u}^*$. In this experiment, we relax the stopping criteria for algorithm 1 to terminate if $|l - l^*| \leqslant 3$ instead of $l = l^*$. The MSE value of the prediction $\tilde{\mathbf{y}} := \mathbf{K}'\mathbf{u}^*$ is defined by $\text{MSE} := \frac{1}{n}\|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2$, and the MSE on the training dataset and the testing dataset are denoted by TrMSE and TeMSE, respectively.

We report in table 5 the targeted sparsity levels l^*, the initial values of λ^0, the selected values of parameter λ^*, the actual sparsity levels SL of the solutions $\mathbf{u}^*$ under the transform $\mathbf{B}$, the numbers NUM of updates for λ^*, the TrMSE values and the TeMSE values. Moreover, we illustrate in figure 4 the convergence process of algorithm 1 for the two prescribed sparsity levels $l^* := 500$ and $l^* := 300$.

6.5. Parameter choices balancing sparsity and accuracy: a separable case

The goal of this subsection is to validate the two parameter choice strategies proposed in theorem 5.2. We consider the signal denoising model (42) with matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ being the Daubechies wavelet transform with $\mathrm{N} := 6$ and $\mathrm{L} := 4$. We again consider recovering the Doppler signal function (49). As in subsection 6.2, we take $n := 4096$ and generate sample points $t_j, j \in \mathbb{N}_n$, the original signal $\mathbf{f}$, and noisy signal $\mathbf{x}^\delta := \mathbf{f} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is the Gaussian noise with noise level $\delta := \|\boldsymbol{\eta}\|_2$.

In the first case, we set seven targeted sparsity levels $l^* = 1819, 1258, 719, 616, 243, 165, 154$. We choose the parameter λ^* according to the strategy stated in item (a) of theorem 5.2. We choose the noise level to be $\delta = 0.0190$. Let $a_j^\delta := \left\|(\mathbf{A}_{(j)})^\top \mathbf{x}^\delta\right\|_\infty$, $j \in \mathbb{N}_n$, rearranged in a

Table 6. Parameter choices $\lambda^* := a^{\delta}_{k_{n-l^*}}$ balancing sparsity level l^* and accuracy ERR for signal denoising (noise $\delta = 0.0190$, total number of wavelet coefficients $n := 4096$).

l^*	1819	1258	719	616	243	165	154
λ^*	2.49×10^{-4}	3.29×10^{-4}	4.60×10^{-4}	4.94×10^{-4}	8.49×10^{-4}	5.28×10^{-3}	7.90×10^{-3}
SL	1819	1258	719	616	243	165	154
ERR	0.0103	0.0092	0.0090	0.0092	0.0135	0.0705	0.1022

Table 7. Parameter choices $\lambda^* := 1.2\delta$ balancing sparsity level SL and accuracy ERR for signal denoising (total number of wavelet coefficients $n := 4096$).

δ	1.90×10^{-9}	7.55×10^{-6}	1.90×10^{-4}	1.90×10^{-2}	6.00×10^{-1}	1.90×10^{0}	6.00×10^{0}
SL	1049	434	274	119	29	15	0
ERR	7.59×10^{-8}	1.94×10^{-4}	3.92×10^{-3}	2.67×10^{-1}	4.73×10^{0}	1.03×10^{1}	1.88×10^{1}
ERR/δ	40.0139	25.7405	20.6788	14.0885	7.8906	5.4288	3.1277

Table 8. Parameter choices $\lambda^* := C\delta$ balancing sparsity level SL and accuracy ERR for signal denoising (total number of wavelet coefficients $n := 4096$).

δ	1.90×10^{-9}				1.90×10^{-4}			
C	0.12	1.2	12	120	0.12	1.2	12	120
SL	1300	1049	833	662	385	274	187	118
ERR	8.50×10^{-9}	7.60×10^{-8}	6.74×10^{-7}	6.01×10^{-6}	4.61×10^{-4}	3.92×10^{-3}	3.24×10^{-2}	2.67×10^{-1}

nondecreasing order: $a^{\delta}_{k_1} \leqslant \cdots \leqslant a^{\delta}_{k_n}$ with distinct $k_i \in \mathbb{N}_n, i \in \mathbb{N}_n$. We choose $\lambda^* := a^{\delta}_{k_{n-l^*}}$ for each l^* and find the corresponding solution $\mathbf{u}^*$ by solving the signal denoising model (42) and determine the actual sparsity level SL of $\mathbf{u}^*$.

We report in table 6 the targeted sparsity levels l^*, the selected values of parameter λ^*, the actual sparsity levels SL of $\mathbf{u}^*$ and the ERR values of $\mathbf{u}^*$. Here and in the next subsection ERR $:= \|\mathbf{u}^* - \mathbf{A}^{-1}\mathbf{f}\|_2$. The SL values in table 6 match exactly with those of the targeted sparsity levels l^*. Moreover, we observe that the ERR values of $\mathbf{u}^*$ depending on λ^* exhibit a pattern like $\lambda^* + 1/\lambda^*$, which is essentially described in item (a) of theorem 5.2 as an upper bound of the approximation error.

In the second case, we choose the parameter according to the second strategy $\lambda^* := C\delta$, described in item (b) of theorem 5.2. We set seven different noise levels δ as shown in the first row of table 7 and choose $C := 1.2$. Numerical results shown in table 7 confirm item (b) of theorem 5.2 which ensures the regularized solution has sparsity of a prescribed level and an approximation error with an upper bound $C'\delta$, where $C' \approx 40.0139$.

We also validate the second parameter choice strategy $\lambda^* := C\delta$, described in item (b) of theorem 5.2 in a different way, by choosing two noise levels $\delta = 1.9 \times 10^{-9}, 1.9 \times 10^{-4}$, and four constants $C = 0.12, 1.2, 12, 120$. The numerical results reported in table 8 are consistent with the theoretical estimate given in item (b) of theorem 5.2.

6.6. Parameter choices balancing sparsity and accuracy: a nonseparable case

The experiments presented in this subsection are to verify the parameter choice strategies described in theorem 5.3. To this end, we consider the signal denoising model (42), where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is determined by the biorthogonal wavelet 'bior2.2' available in Matlab with the decomposition levels L := 6. Here, 'bior2.2' denotes the two wavelets for decomposition and

Table 9. Parameter choices λ^* balancing sparsity level l^* and accuracy ERR for signal denoising (noise $\delta = 0.0190$, initial parameter $\lambda^0 = 3.9478$ and total number of wavelet coefficients $n := 4096$).

l^*	600	1000	1600	2200	2800	3400	3600
λ^*	1.02×10^{-3}	5.57×10^{-4}	3.62×10^{-4}	2.48×10^{-4}	1.58×10^{-4}	8.26×10^{-5}	5.90×10^{-5}
SL	600	1000	1600	2200	2800	3400	3600
NUM	6	5	8	5	6	5	8
ERR	0.0505	0.0319	0.0239	0.0198	0.0179	0.0180	0.0185

Table 10. Parameter choices $\lambda^* := 1.2\delta$ balancing sparsity level SL and accuracy ERR for signal denoising (total number of wavelet coefficients $n := 4096$).

δ	2.39×10^{-5}	1.90×10^{-4}	1.07×10^{-3}	1.90×10^{-2}	6.00×10^{-1}	1.90×10^{0}	6.00×10^{0}
SL	1626	920	531	167	51	24	0
ERR	2.45×10^{-3}	1.47×10^{-2}	6.05×10^{-2}	5.23×10^{-1}	7.06×10^{0}	1.57×10^{1}	1.96×10^{1}
ERR$/\delta$	102.4764	77.3039	56.7137	27.5954	11.7829	8.2608	3.2647

Table 11. Parameter choices $\lambda^* := C\delta$ balancing sparsity level SL and accuracy ERR for signal denoising (total number of wavelet coefficients $n := 4096$).

δ	1.90×10^{-4}				1.90×10^{-2}			
C	0.12	1.2	12	120	0.12	1.2	12	120
SL	1712	920	422	167	424	167	79	24
ERR	2.00×10^{-3}	1.47×10^{-2}	9.52×10^{-2}	5.23×10^{-1}	9.51×10^{-2}	5.23×10^{-1}	2.97×10^{0}	1.57×10^{1}

reconstruction respectively having the same order N := 2 of vanishing moments. In this case, **A** is not separable. We again recover the Doppler signal function (49) from the noisy signal $\mathbf{x}^\delta$ defined as in subsection 6.5.

We first set seven targeted sparsity levels l^*, which are $l^* = 600, 1000, 1600, 2200, 2800, 3400, 3600$. According to item (a) of theorem 5.3, we choose the parameter λ^* by employing algorithm 1 with $\psi(\mathbf{u}) := \frac{1}{2}\left\|\mathbf{A}\mathbf{u} - \mathbf{x}^\delta\right\|_2^2$, $\mathbf{B} := \mathbf{I}_n$ and $\mathbf{B}' = \mathbf{I}_n$. Numerically, we choose the initial parameter $\lambda^0 = \left\|\mathbf{A}^\top \mathbf{x}^\delta\right\|_\infty$ ($\approx$3.9478). The targeted sparsity levels l^*, the selected values of parameter λ^* chosen by algorithm 1, the actual sparsity levels SL of the solutions $\mathbf{u}^*$, the numbers NUM of updates for λ^* and the ERR values of the solutions $\mathbf{u}^*$ are reported in table 9. The ERR values of $\mathbf{u}^*$ depending on λ^* exhibit a pattern like $\lambda^* + 1/\lambda^*$, which is essentially described in item (a) of theorem 5.3 as an upper bound of the approximation error.

We next set seven different noise levels δ as shown in the first row of table 10 and choose the values of the parameter λ^* according to $\lambda^* := C\delta$ with $C = 1.2$. The numerical results for this case are reported in table 10, where the noise levels δ, the actual sparsity levels SL of the solutions $\mathbf{u}^*$, the ERR values of $\mathbf{u}^*$ and ERR$/\delta$ are listed.

For the parameter choice strategy $\lambda^* := C\delta$ proposed in item (b) of theorem 5.3, we consider two values $\delta = 1.9 \times 10^{-4}, 1.9 \times 10^{-2}$, and four values $C = 0.12, 1.2, 12, 120$. Numerical results are reported in table 11. The numerical results in both tables 10 and 11 confirm the theoretical results in item (b) of theorem 5.3.

Table 12. ℓ_1 SVM classification model with hinge loss (512 training dataset).

λ^*	0.1	0.2	1	2	4	10	27.9851
γ	0.0543	0.1458	0.8785	1.9399	3.7844	9.6159	27.9850
SL	151	151	111	56	37	15	0
TrA	100%	100%	97.27%	95.51%	92.97%	81.45%	50.00%
TeA	96.71%	96.66%	95.68%	93.47%	91.36%	80.85%	50.47%

6.7. ℓ_1 SVM classification with the hinge loss function and regression with the ϵ-insensitive loss function

This example tests the result in theorem 4.5 by considering the ℓ_1 SVM classification/regression models (31) and (33). Theorem 4.5 applied to these two models leads to corollary 4.7 and corollary 4.8, respectively. Note that for these models, the choice of parameter λ^* in Corollaries 4.7 and 4.8 depends on the unknown solution $\mathbf{u}^*$. Hence, we test the necessary condition described by the inequalities in (32) of corollary 4.7 and the inequalities in (34) of corollary 4.8 for problems (31) and (33), respectively, having a solution $\mathbf{u}^*$ with a prescribed sparsity level under the transform $\mathbf{B}$. Specifically, for a given parameter λ^*, by solving problems (31) and (33), we find the corresponding solutions $\mathbf{u}^*$, and then verify if the pair of the chosen λ^* value and the corresponding solution $\mathbf{u}^*$ satisfy the inequalities in (32), and if the sparsity level of $\mathbf{Bu}^*$ matches the one described in corollary 4.7. Likewise, we conduct the same test for corollary 4.8.

We first consider the ℓ_1 SVM classification model (31) with the hinge loss function. The dataset we use is MNIST database with digits '7' and '9' as mentioned in subsection 6.4. The kernel we choose for (31) is the Gaussian kernel defined by (50) with $\mu := 4$ and $d := 784$. Let n be the number of training samples and $\mathbf{y} \in \{-1, 1\}^n$ the given vector storing the labels of the training data, where -1 and 1 represent the digits '7' and '9' respectively. The experiment uses 512 training samples and 2037 testing samples of these two digits from the database. We choose seven different values of parameter λ^* listed in the first row of table 12.

Associated with each solution $\mathbf{u}^*$, we identify l distinct integers $k_i \in \mathbb{N}_n$, $i \in \mathbb{N}_l$, so that $\mathbf{Bu}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i}$, $z^*_{k_i} \in \mathbb{R}\setminus\{0\}$, $i \in \mathbb{N}_l$. That is, $\mathbf{Bu}^*$ has l nonzero components. We compute the number

$$\gamma := \max\left\{\min\left\{|(\mathbf{YK}_j)^\top \mathbf{c}| : \mathbf{c} \in \partial\phi(\mathbf{YK}'\mathbf{u}^*)\right\} : j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}\right\}$$

and verify indeed that $\lambda^* \geqslant \gamma$. The values of γ, the actual sparsity levels SL of $\mathbf{Bu}^*$, the TrA values and the TeA values are reported in table 12. These numerical results confirm the inequality in (32) of corollary 4.7. Note that if λ^* is sufficiently large, the vector $\mathbf{Bu}^*$ has the sparsity level 0. The value of the parameter λ^* listed in the last column of table 12 produces a solution that has most sparsity under the transform $\mathbf{B}$.

We repeat the experiment described above with the same kernel by using 8141 training samples and 2037 testing samples. The selected values of parameter λ^*, the actual sparsity levels SL of $\mathbf{Bu}^*$, the TrA values and the TeA values are reported in table 13. Observing form table 13, we get that the sparsity level of $\mathbf{Bu}^*$ is smaller as the value of parameter λ^* increases, the corresponding TrA and TeA values become lower.

Secondly, we consider the ℓ_1 SVM regression model (33) with the ϵ-insensitive loss function. The dataset we use is 'Mg' as mentioned in subsection 6.4. The kernel used is the Gaussian kernel (50) with $\mu := 1.5$ and $d := 6$. Let $n := 1000$ and $\mathbf{y} \in \mathbb{R}^n$ be the given labels of the training data. The parameter ϵ involved in the ϵ-insensitive loss for model (33) is chosen

Table 13. ℓ_1 SVM classification model with hinge loss (8141 training dataset).

λ^*	0.1	0.2	1	2	4	10	435.0694
SL	552	481	167	92	56	34	0
TrA	99.99%	99.99%	99.08%	98.17%	97.53%	96.30%	50.67%
TeA	98.72%	98.77%	98.38%	98.09%	97.45%	96.27%	50.47%

Table 14. ℓ_1 SVM regression model with ϵ-insensitive loss (1000 training dataset).

λ^*	0.01	0.4	1	2	4	18.00	135.8091
γ	0.005	0.1045	0.8980	1.7528	3.2839	17.8510	135.8091
SL	305	33	19	11	7	5	0
TrMSE	0.0145	0.0165	0.0177	0.0185	0.0200	0.0207	0.0530
TeMSE	0.0157	0.0162	0.0170	0.0177	0.0193	0.0202	0.0530

Table 15. Total-variation signal denoising model (total sparsity level $n := 4095$).

λ^*	0.1	0.2	0.5	6	30	107	270.1717
SL	1488	744	355	160	67	12	0
l	1513	759	361	164	67	12	1
MSE	0.0042	0.0020	0.0012	0.0076	0.0297	0.0656	0.0835

as $\epsilon := 10^{-4}$. We choose seven different values of parameter λ^* as shown in the first row of table 14.

For each solution $\mathbf{u}^*$, we identify l distinct integers $k_i \in \mathbb{N}_n$, $i \in \mathbb{N}_l$, so that $\mathbf{B}\mathbf{u}^* := \sum_{i\in\mathbb{N}_l} z^*_{k_i}\mathbf{e}_{k_i}$, $z^*_{k_i} \in \mathbb{R}\setminus\{0\}$, $i \in \mathbb{N}_l$. That is, $\mathbf{B}\mathbf{u}^*$ has l nonzero components. We compute the number $\gamma := \max\left\{\min\left\{\left|(\mathbf{K}_j)^\top\mathbf{c}\right| : \mathbf{c} \in \partial\phi_{\mathbf{y},\epsilon}(\mathbf{K}'\mathbf{u}^*)\right\} : j \in \mathbb{N}_n \setminus \{k_i : i \in \mathbb{N}_l\}\right\}$. The selected values of parameter λ^*, the values of γ, the actual sparsity levels SL of $\mathbf{B}\mathbf{u}^*$, the TrMSE values and the TeMSE values are reported in table 14. The numerical results reported in table 14 confirm the inequalities in (34) of corollary 4.8.

6.8. *Total-variation signal denoising*

In this experiment, we verify the result in proposition 4.15. We again consider recovering the Doppler signal function defined by (49) from its noisy data by the total-variation signal denoising model (35). The original signal $\mathbf{f}$ and the noisy signal $\mathbf{x}$ are chosen in the same way as in subsection 6.2. Note that the fidelity term $\psi(\mathbf{u}) := \frac{1}{2}\left\|\mathbf{u}-\mathbf{x}\right\|_2^2$, $\mathbf{u} \in \mathbb{R}^n$, is differentiable and the transform matrix $\mathbf{B} := \mathbf{D}^{(1)}$, an $m \times n$ matrix, has full row rank. In this case, $m = n-1$. According to remark 4.10, when $\lambda \geqslant \lambda_{\max} := \left\|(\widetilde{\mathbf{D}}^{(1)\prime})^\top\mathbf{x}\right\|_\infty$ ($\approx$270.1717), the corresponding regularized solution is the zero vector under the transform $\mathbf{D}^{(1)}$. For this reason, we choose seven different values of the parameter λ^* in the interval $(0, \lambda_{\max}]$. For each selected value of λ^*, we find the corresponding solution $\mathbf{u}^*$ by solving model (35) using FPPA and count the actual sparsity level of $\mathbf{D}^{(1)}\mathbf{u}^*$. To verify the result in proposition 4.15, we choose $\epsilon := 5\times10^{-5}$ and $\mathbf{v} := \mathbf{u}^* + \epsilon\mathbf{v}_0/\|\mathbf{v}_0\|_2$, where $\mathbf{v}_0 \in [0,1]^n$ is a uniformly distributed random vector. We then use the vector $\mathbf{v}$ that satisfies $\|\mathbf{u}^*-\mathbf{v}\|_2 \leqslant \epsilon$ to define the sequence $v^\epsilon_j := \left|(\mathbf{B}'_j)^\top(\mathbf{v}-\mathbf{x})\right| + \epsilon L_j$ with $L_j := \|\mathbf{B}'_j\|_2, j \in \mathbb{N}_m$. We rearrange the sequence in a nondecreasing order: $v^\epsilon_{k_1} \leqslant \cdots \leqslant v^\epsilon_{k_m}$ with distinct $k_i \in \mathbb{N}_m$, for $i \in \mathbb{N}_m$, which allows us to identify the integer $l \in \mathbb{Z}_{m+1}$ such that $v^\epsilon_{k_1} \leqslant \cdots \leqslant v^\epsilon_{k_{m-l}} < \lambda^* \leqslant v^\epsilon_{k_{m-l+1}} \leqslant \cdots \leqslant v^\epsilon_{k_m}$, that is, condition (39) is satisfied.

We report in table 15 the selected values of parameter λ^*, the actual sparsity levels SL of $\mathbf{D}^{(1)}\mathbf{u}^*$, the integer l determined by proposition 4.15 and the MSE values of the denoised

signals $\mathbf{u}^*$, where the MSE is defined as in subsection 6.3. These numerical results confirm that the actual sparsity level of $\mathbf{D}^{(1)}\mathbf{u}^*$, which corresponds to each selected value of λ^*, does not exceed l determined by inequality (39).

To close this section, we remark that both the proposed parameter choice strategy and the iterative algorithm perform well in overcoming overfitting when data are contaminated with noise, while producing solutions with desired sparsity levels.

7. Conclusion

We have studied choice strategies of the regularization parameter for various regularization problems with an ℓ_1 norm regularization. The strategies are proposed to balance sparsity of a regularized solution and its approximation accuracy compared to the corresponding minimal norm solution. The ingredient used in developing the strategies is the connection of the choice of the parameter with the sparsity level of the regularized solution and with the approximation error. Much effort of this paper is given to understanding the connection between the choice of the parameter and the sparsity level of the regularized solution. We have also demonstrated how such understanding is combined with an error bound of the regularized solution to obtain a strategy for choices of the parameter to balance its sparsity and approximation accuracy. We have conducted substantial numerical experiments to test the proposed strategies. Numerical results of various application models confirm our theoretical estimates. More extensive applications of the proposed strategies and convergence analysis of the proposed algorithm for choosing the parameter will be our future research projects.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

R Wang is supported in part by the National Key Research and Development Program of China (Grant Nos. 2020YFA0714100 and 2020YFA0713600) and by the Natural Science Foundation of China under Grant 12171202; Y Xu is supported in part by the US National Science Foundation under Grants DMS-1912958 and DMS-2208386, and by the US National Institutes of Health under Grant R21CA263876. The authors are grateful to anonymous referees for their constructive comments which have improved the quality of this paper.

Appendix

Proof of lemma 4.1

Substituting $\mathbf{B}^\dagger = \mathbf{V}\mathbf{\Lambda}^\dagger\mathbf{U}^\top$ into the representation of the general solution $\mathbf{u}$ of (20), we have that $\mathbf{u} = \mathbf{V}\left(\mathbf{\Lambda}^\dagger\mathbf{U}^\top\mathbf{z} + \begin{bmatrix}\mathbf{0}\\ \mathbf{v}\end{bmatrix}\right)$ for any $\mathbf{v} \in \mathbb{R}^{n-r}$. In terms of $\mathbf{\Lambda}'$ and $\mathbf{U}'$, the above equation can be rewritten as $\mathbf{u} = \mathbf{V}\mathbf{\Lambda}'\mathbf{U}'\begin{bmatrix}\mathbf{z}\\ \mathbf{v}\end{bmatrix}$. By using $\mathbf{B}'$, the above equation can be represented as the desired form (21).

It remains to verify that for each solution $\mathbf{u}$ of (20), the vector $\mathbf{v}$ appearing in (21) is unique. Suppose that $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{n-r}$ both satisfy (21). We then obtain that $\mathbf{B}'\begin{bmatrix}\mathbf{0}\\ \mathbf{v}_1 - \mathbf{v}_2\end{bmatrix} = \mathbf{0}$, which

together with the definition of $\mathbf{B}'$ implies that $\mathbf{V}\begin{bmatrix} 0 \\ \mathbf{v}_1 - \mathbf{v}_2 \end{bmatrix} = \mathbf{0}$. The invertibility of $\mathbf{V}$ ensures that $\mathbf{v}_1 = \mathbf{v}_2$, proving the desired result.

Proof of lemma 4.2

It suffices to show that $\mathcal{B}$ is surjective and injective. We first verify the surjectivity of $\mathcal{B}$. For any $\mathbf{z} \in \mathcal{R}(\mathbf{B})$ and any $\mathbf{v} \in \mathbb{R}^{n-r}$, we define a vector $\mathbf{u} \in \mathbb{R}^n$ through (21). Lemma 4.1 guarantees that $\mathbf{Bu} = \mathbf{z}$, which together with the definition of $\mathcal{B}$ implies that $\mathcal{B}\mathbf{u} = \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix}$. Hence, $\mathcal{B}$ is surjective. To prove the injectivity of $\mathcal{B}$, we suppose that $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$ satisfy $\mathcal{B}\mathbf{u}_1 = \mathcal{B}\mathbf{u}_2$. It follows from $\mathbf{B}'\mathcal{B}\mathbf{u}_1 = \mathbf{u}_1$ and $\mathbf{B}'\mathcal{B}\mathbf{u}_2 = \mathbf{u}_2$ that $\mathbf{u}_1 = \mathbf{B}'\mathcal{B}\mathbf{u}_1 = \mathbf{B}'\mathcal{B}\mathbf{u}_2 = \mathbf{u}_2$. That is, $\mathcal{B}$ is injective.

Proof of proposition 4.3

Note that the mapping $\mathcal{B}$ provides a bijective correspondence between $\mathbb{R}^n$ and $\mathcal{R}(\mathbf{B}) \times \mathbb{R}^{n-r}$. It suffices to verify that for all $\mathbf{u} \in \mathbb{R}^n$ there holds $\psi(\mathbf{u}) + \lambda\|\mathbf{Bu}\|_1 = \psi \circ \mathbf{B}'(\mathcal{B}\mathbf{u}) + \lambda\|\mathbf{I}'\mathcal{B}\mathbf{u}\|_1$. By the definition of $\mathcal{B}$, we have that $\mathbf{I}'\mathcal{B}\mathbf{u} = \mathbf{Bu}$. This together with $\mathbf{B}'\mathcal{B}\mathbf{u} = \mathbf{u}$ confirms the validity of the equation above.

Proof of remark 4.10

It follows from corollary 4.9 with $l = 0$ that the total-variation signal denoising model has a solution $\mathbf{u}^*$ satisfying $\mathbf{D}^{(1)}\mathbf{u}^* = 0$ if and only if there hold

$$\lambda \geqslant \left\|(\widetilde{\mathbf{D}}^{(1)'})^\top(\mathbf{u}^* - \mathbf{x})\right\|_\infty \text{ and } 1_n^\top(\mathbf{u}^* - \mathbf{x}) = 0. \tag{51}$$

Suppose that $\mathbf{D}^{(1)}\mathbf{u}^* = \mathbf{0}$. We obtain $\mathbf{u}^*$ by solving two equations $\mathbf{D}^{(1)}\mathbf{u}^* = \mathbf{0}$ and $1_n^\top(\mathbf{u}^* - \mathbf{x}) = 0$. By lemma 4.1, the vector $\mathbf{u}^*$ satisfying the first equation can be represented as $\mathbf{u}^* = \mathbf{D}^{(1)'}\begin{bmatrix} \mathbf{0} \\ v^* \end{bmatrix}$ for some $v^* \in \mathbb{R}$, which together with $\mathbf{D}_n^{(1)'} = \frac{\sqrt{n}}{n}1_n$, leads to $\mathbf{u}^* = \frac{v^*\sqrt{n}}{n}1_n$. Substituting this representation into the second equation yields that $v^* = \frac{\sqrt{n}}{n}1_n^\top\mathbf{x}$, which further leads to $\mathbf{u}^* = \frac{1}{n}(1_n^\top\mathbf{x})1_n$. This allows us to rewrite the inequality in condition (51) as $\lambda \geqslant \left\|\frac{1}{n}(1_n^\top\mathbf{x})(\widetilde{\mathbf{D}}^{(1)'})^\top 1_n - (\widetilde{\mathbf{D}}^{(1)'})^\top\mathbf{x}\right\|_\infty$. It suffices to show $(\widetilde{\mathbf{D}}^{(1)'})^\top 1_n = \mathbf{0}$. By the definition of $\mathbf{D}^{(1)'}$, we have $(\mathbf{D}^{(1)'})^\top\mathbf{V}_n = (\mathbf{U}')^\top\mathbf{\Lambda}'\mathbf{V}^\top\mathbf{V}_n$. Substituting $\mathbf{V}^\top\mathbf{V}_n = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$ into the above equation, we get $(\mathbf{D}^{(1)'})^\top\mathbf{V}_n = (\mathbf{U}')^\top\mathbf{\Lambda}'\begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$, which further yields $(\mathbf{D}^{(1)'})^\top\mathbf{V}_n = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$. That is, $(\widetilde{\mathbf{D}}^{(1)'})^\top\mathbf{V}_n = \mathbf{0}$. Noting $\mathbf{V}_n = \frac{\sqrt{n}}{n}1_n$, we obtain $(\widetilde{\mathbf{D}}^{(1)'})^\top 1_n = \mathbf{0}$. Thus, we rewrite the inequality in condition (51) as $\lambda \geqslant \left\|(\widetilde{\mathbf{D}}^{(1)'})^\top\mathbf{x}\right\|_\infty$. Conversely, suppose that $\lambda \geqslant \left\|(\widetilde{\mathbf{D}}^{(1)'})^\top\mathbf{x}\right\|_\infty$. By setting $\mathbf{u}^* := \frac{1}{n}(1_n^\top\mathbf{x})1_n$, we conclude that condition (51) holds. That is, $\mathbf{u}^*$ is a solution of the total-variation signal denoising model with $\mathbf{D}^{(1)}\mathbf{u}^* = \mathbf{0}$.

ORCID iD

Yuesheng Xu https://orcid.org/0000-0003-2982-7864

References

[1] Ali A and Tibshirani R J 2019 The generalized lasso problem and uniqueness *Electron. J. Stat.* **13** 2307–47

[2] Anzengruber S W and Ramlau R 2010 Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators *Inverse Problems* **26** 025001

[3] Argyriou A, Micchelli C A, Pontil M, Shen L and Xu Y 2011 Efficient first order methods for linear composite regularizers (arXiv:1104.1436)

[4] Bach F, Jenatton R, Mairal J and Obozinski G 2012 Optimization with sparsity-inducing penalties *Found. Trends Mach. Learn.* **4** 1–106

[5] Bauer F and Lukas M A 2011 Comparing parameter choice methods for regularization of ill-posed problems *Math. Comput. Simul.* **81** 1795–841

[6] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202

[7] Bi J, Bennett K P, Embrechts M, Breneman C M and Song M 2003 Dimensionality reduction via sparse support vector machines *J. Mach. Learn. Res.* **3** 1229–43

[8] Björck Å 1996 *Numerical Methods for Least Squares Problems* (Philadelphia, PA: SIAM)

[9] Bonesky T 2009 Morozov's discrepancy principle and Tikhonov-type functionals *Inverse Problems* **25** 015015

[10] Boyd S, Parikh N, Chu E, Peleato B and Eckstein J 2011 Distributed optimization and statistical learning via the alternating direction method of multipliers *Found. Trends Mach. Learn.* **3** 1–122

[11] Bühlmann P and Van De Geer S 2011 *Statistics for High-Dimensional Data: Methods, Theory and Applications* (New York: Springer)

[12] Cai J-F, Candés E J and Shen Z 2010 A singular value thresholding algorithm for matrix completion *SIAM J. Optim.* **20** 1956–82

[13] Chang C-C and Lin C-J 2011 Libsvm: a library for support vector machines *ACM Trans. Intell. Syst. Technol.* **2** 1–27

[14] Chang S G, Yu B and Vetterli M 2000 Adaptive wavelet thresholding for image denoising and compression *IEEE Trans. Image Process.* **9** 1532–46

[15] Chen S S, Donoho D L and Saunders M A 2001 Atomic decomposition by basis pursuit *SIAM Rev.* **43** 129–59

[16] Chen Z, Micchelli C M and Xu Y 2015 *Multiscale Methods for Fredholm Integral Equations* (Cambridge: Cambridge University Press)

[17] Cheng R and Xu Y 2021 Minimum norm interpolation in the ℓ_1 space *Anal. Appl.* **19** 21–42

[18] Condat L 2013 A direct algorithm for 1D total variation denoising *IEEE Signal Process. Lett.* **20** 1054–7

[19] Cox D D and O'Sullivan F 1990 Asymptotic analysis of penalized likelihood and related estimators *Ann. Stat.* **18** 1676–95

[20] Daubechies I 1992 *Ten Lectures on Wavelets* (Philadelphia, PA: SIAM)

[21] de Prado M L 2020 Overfitting: causes and solutions (seminar slides) *SSRN 3544431* (https://doi.org/10.2139/ssrn.3544431)V

[22] Donoho D L and Johnstone I M 1995 Adapting to unknown smoothness via wavelet shrinkage *J. Am. Stat. Assoc.* **90** 1200–24

[23] Egger H and Hofmann B 2018 Tikhonov regularization in Hilbert scales under conditional stability assumptions *Inverse Problems* **34** 115015

[24] Gasquet C and Witomski P 1999 *Fourier Analysis and Applications* (New York: Springer)

[25] Gilbert J C 2017 On the solution uniqueness characterization in the L1 norm and polyhedral gauge recovery *J. Optim. Theory Appl.* **172** 70–101

[26] Grasmair M, Haltmeier M and Scherzer O 2008 Sparse regularization with l^q penalty term *Inverse Problems* **24** 055020

[27] Horn R A and Johnson C R 1985 *Matrix Analysis* (Cambridge: Cambridge University Press)

[28] Jafarpour B, Goyal V K, McLaughlin D B and Freeman W T 2009 Transform-domain sparsity regularization for inverse problems in geosciences *Geophysics* **74** R69–R83

[29] Jenatton R, Mairal J, Obozinski G and Bach F 2011 Proximal methods for hierarchical sparse coding *J. Mach. Learn. Res.* **12** 2297–334

[30] Juditsky A, Karzan F K, Nemirovski A and Polyak B 2012 Accuracy guaranties for ℓ_1 recovery of block-sparse signals *Ann. Stat.* **40** 3077–107

[31] Kim S-J, Koh K, Boyd S and Gorinevsky D 2009 ℓ_1 trend filtering *SIAM Rev.* **51** 339–60
[32] Kimeldorf G S and Wahba G 1970 A correspondence between Bayesian estimation on stochastic processes and smoothing by splines *Ann. Math. Stat.* **41** 495–502
[33] Koh K, Kim S J and Boyd S 2007 An interior-point method for large-scale ℓ_1-regularized logistic regression *J. Mach. Learn. Res.* **8** 1519–55
[34] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324
[35] Li Q, Shen L, Xu Y and Zhang N 2015 Multi-step fixed-point proximity algorithms for solving a class of optimization problems arising from image processing *Adv. Comput. Math.* **41** 387–422
[36] Li Y-R, Shen L, Dai D-Q and Suter B W 2011 Framelet algorithms for de-blurring images corrupted by impulse plus Gaussian noise *IEEE Trans. Image Process.* **20** 1822–37
[37] Li Z, Song G and Xu Y 2018 A fixed-point proximity approach to solving the support vector regression with the grouplasso regularization *Int. J. Numer. Anal. Model.* **15** 154–69
[38] Li Z, Song G and Xu Y 2019 A two-step fixed-point proximity algorithm for a class of non-differentiable optimization models in machine learning *J. Sci. Comput.* **81** 923–40
[39] Lian Q, Shen L, Xu Y and Yang L 2011 Filters of wavelets on invariant sets for image denoising *Appl. Anal.* **90** 1299–322
[40] Lin R, Song G and Zhang H 2021 Multi-task learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm *J. Complexity* **63** 101514
[41] Lorenz D A 2008 Convergence rates and source conditions for Tikhonov regularization with sparsity constraints *J. Inverse Ill-Posed Problems* **16** 463–78
[42] McCann M T and Ravishankar S 2020 Supervised learning of sparsity-promoting regularizers for denoising (arXiv:2006.05521)
[43] Micchelli C A, Shen L and Xu Y 2011 Proximity algorithms for image models: denoising *Inverse Problems* **27** 045009
[44] Micchelli C A and Xu Y 1997 Reconstruction and decomposition algorithms for biorthogonal multiwavelets *Multidimens. Syst. Signal Process.* **8** 31–69
[45] Morozov V A 1966 On the solution of functional equations by the method of regularization *Sov. Math. Dokl.* **7** 414–7
[46] Pereverzev S and Schock E 2005 On the adaptive selection of the parameter in regularization of ill-posed problems *SIAM J. Numer. Anal.* **43** 2060–76
[47] Jin Q N 1999 Applications of the modified discrepancy principle to Tikhonov regularization of nonlinear ill-posed problems *SIAM J. Numer. Anal.* **36** 475–90
[48] Ramlau R and Teschke G 2006 A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints *Numer. Math.* **104** 177–203
[49] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica* D **60** 259–68
[50] Schölkopf B, Herbrich R and Smola A J 2001 A generalized representer theorem *Proc. 14th Annual Conf. on Computational Learning Theory and the Fifth European Conf. on Computational Learning Theory* (London: Springer) pp 416–26
[51] Schölkopf B and Smola A J 2002 *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond* (Cambridge: MIT Press)
[52] Schuster T, Kaltenbacher B, Hofmann B and Kazimierski K S 2012 *Regularization Methods in Banach Spaces* (Berlin: de Gruyter)
[53] Shepard N G 1990 The singular-value decomposition of the first-order difference matrix *Econ. Theory* **6** 119–20
[54] Shi L, Feng Y and Zhou D 2011 Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces *Appl. Comput. Harmon. Anal.* **31** 286–302
[55] Smola A J and Schölkopf B 2004 A tutorial on support vector regression *Stat. Comput.* **14** 199–222
[56] Song G and Zhang H 2011 Reproducing kernel Banach spaces with the ℓ^1 norm II: error analysis for regularized least square regression *Neural Comput.* **23** 2713–29
[57] Song G, Zhang H and Hickernell F J 2013 Reproducing kernel Banach spaces with the ℓ^1 norm *Appl. Comput. Harmon. Anal.* **34** 96–116
[58] Strang G 1999 The discrete cosine transform *SIAM Rev.* **41** 135–47
[59] Tautenhahn U 1996 Error estimates for regularization methods in Hilbert scales *SIAM J. Numer. Anal.* **33** 2120–30
[60] Tautenhahn U 1998 On a general regularization scheme for nonlinear ill-posed problems: II. Regularization in Hilbert scales *Inverse Problems* **14** 1607–16

[61] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267–88
[62] Tibshirani R, Saunders M, Rosset S, Zhu J and Knight K 2005 Sparsity and smoothness via the fused lasso *J. R. Stat. Soc.* B **67** 91–108
[63] Tibshirani R J and Taylor J 2011 The solution path of the generalized lasso *Ann. Stat.* **39** 1335–71
[64] Tibshirani R J and Taylor J 2012 Degrees of freedom in lasso problems *Ann. Stat.* **40** 1198–232
[65] Tikhonov A and Arsenin V 1977 *Solutions to Ill-Posed Problems* (New York: Wiley)
[66] Tomassi D, Milone D and Nelson J D B 2015 Wavelet shrinkage using adaptive structured sparsity constraints *Signal Process.* **106** 73–87
[67] Unser M 2021 A unifying representer theorem for inverse problems and machine learning *Found. Comput. Math.* **21** 941–60
[68] Vapnik V N 1998 *Statistical Learning Theory* (New York: Wiley)
[69] Wang R and Xu Y 2021 Representer theorems in Banach spaces: minimum norm interpolation, regularized learning and semi-discrete inverse problems *J. Mach. Learn. Res.* **22** 1–65
[70] Wang Y, Sharpnack J, Smola A J and Tibshirani R J 2016 Trend filtering on graphs *J. Mach. Learn. Res.* **17** 1–41
[71] Xu Y 2022 Sparse regularization with the ℓ_0 norm *Anal. Appl.* 1–29
[72] Xu Y and Ye Q 2019 *Generalized Mercer Kernels and Reproducing Kernel Banach Spaces* vol 258 (Providence, RI: *Memoirs of the American Mathematical Society*) p 1243
[73] Yang J and Zhang Y 2011 Alternating direction algorithms for ℓ_1-problems in compressive sensing *SIAM J. Sci. Comput.* **33** 250–78
[74] Yuan M and Lin Y 2006 Model selection and estimation in regression with grouped variables *J. R. Stat. Soc.* B **68** 49–67
[75] Zălinescu C 2002 *Convex Analysis in General Vector Spaces* (River Edge, NJ: World Scientific)
[76] Zhang H, Xu Y and Zhang J 2009 Reproducing kernel Banach spaces for machine learning *J. Mach. Learn. Res.* **10** 2741–75
[77] Zou H, Hastie T and Tibshirani R 2007 On the "degrees of freedom" of the lasso *Ann. Stat.* **35** 2173–92