Bin Zhu; Yijie Shi

MKL-$L_{0/1}$-SVM

Kybernetika, Vol. 62 (2026), No. 3, 481–512

Persistent URL: http://dml.cz/dmlcz/153684

MKL-$L_{0/1}$-SVM

Bin Zhu and Yijie Shi

This paper presents a Multiple Kernel Learning (abbreviated as MKL) framework for the Support Vector Machine (SVM) with the $(0,1)$ loss function in the context of the binary classification task. Some KKT-like first-order optimality conditions are provided and then exploited to develop a fast ADMM algorithm to solve the nonsmooth nonconvex optimization problem. Convergence analysis of the ADMM is presented under certain technical assumptions. Numerical experiments on real data sets show that our MKL-$L_{0/1}$-SVM can potentially outperform one of the leading approaches called SimpleMKL developed by Rakotomamonjy, Bach, Canu, and Grandvalet [Journal of Machine Learning Research, vol. 9, pp. 2491–2521, 2008] in terms of the accuracy of classification and the sparsity in the combination of kernels.

Keywords: kernel SVM, (0,1)-loss function, nonsmooth nonconvex optimization, Multiple Kernel Learning, alternating direction method of multipliers

Classification: 46E22, 90C26

1. INTRODUCTION

The support vector machine (SVM) is an important tool in machine learning with numerous applications [21, 22]. The theory can be traced back to the seminal work [8]. In the basic setting, the SVM deals with the binary classification task in which a data set $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i = 1, \ldots, m\}$ of size $m \in \mathbb{N}_+$ is given and it is asked to construct a function $\tilde{f}(\mathbf{x})$ in order to separate the two classes of feature vectors $\mathbf{x}_i$ with labels $y_i = -1$ or 1, and to predict the labels of unseen feature vectors. To this end, the SVM first lifts the problem to a *reproducing kernel Hilbert space*[1] (RKHS) $\mathbb{H}$, in general infinite-dimensional and equipped with a *positive definite* kernel function $\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, via the feature mapping

$$\mathbf{x} \mapsto \phi(\mathbf{x}) := \kappa(\cdot, \mathbf{x}) \in \mathbb{H}, \tag{1}$$

and then considers decision (or discriminant) functions of the form

$$\tilde{f}(\mathbf{x}) = b + \langle w, \phi(\mathbf{x}) \rangle_{\mathbb{H}} = b + w(\mathbf{x}), \tag{2}$$

where $b \in \mathbb{R}$, $w \in \mathbb{H}$, $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ the inner product associated to the RKHS $\mathbb{H}$, and the second equality is due to the so-called *reproducing property*. It is not difficult to see

DOI: 10.14736/kyb-2026-3-0481

[1] The theory of RKHS goes back to the classic work [1] and many more, see e. g., [15].

that such a decision function is in general nonlinear in $\mathbf{x}$, but is indeed linear with respect to $\phi(\mathbf{x})$ in the feature space $\mathbb{H}$. Once $\tilde{f}$ is determined, the label of $\mathbf{x}$ is assigned via $y(\mathbf{x}) = \text{sign}[\tilde{f}(\mathbf{x})]$ where sign$(\cdot)$ is the sign function which gives $+1$ for a positive number, -1 for a negative number, and is left undefined at zero.

It remains to estimate the unknown quantities b and w in (2). When the data $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ are *linearly separable* in the RKHS sense, the parameter estimation is formulated as an optimization problem

$$\min_{\substack{w\in\mathbb{H},\, b\in\mathbb{R},\\ \tilde{f}(\cdot)=w(\cdot)+b}} \quad \frac{1}{2}\|w\|_{\mathbb{H}}^2 \quad \text{s.t.} \quad y_i\tilde{f}(\mathbf{x}_i) \geq 1, \quad i = 1, \ldots, m, \tag{3}$$

where $\|w\|_{\mathbb{H}}^2 = \langle w, w\rangle_{\mathbb{H}}$ is the squared norm of w induced by the inner product. The separability condition is described by the set of inequalities in (3) and the objective function is inversely proportional to the margin between the decision hyperplane in $\mathbb{H}$ and the nearest points in each class. Thus, the SVM is a maximum-margin classifier subject to the separability constraint. A solution to the optimization problem (3) is usually called a *hard-margin SVM*.

In the more general case with *linearly inseparable* data, which is the focus of consideration in this paper, a suitable loss function $\mathcal{L}(\cdot,\cdot)$ must be employed to quantify the error of classification since misclassification is inevitable. The resulting classifier is called a *soft-margin SVM* which solves the unconstrained optimization problem

$$\min_{\substack{w\in\mathbb{H},\, b\in\mathbb{R},\\ \tilde{f}(\cdot)=w(\cdot)+b}} \quad \frac{1}{2}\|w\|_{\mathbb{H}}^2 + C\sum_{i=1}^m \mathcal{L}(y_i, \tilde{f}(\mathbf{x}_i)), \tag{4}$$

where $C > 0$ is a regularization parameter. For the choice of $\mathcal{L}$, the original paper [8] on SVM suggested the $(0,1)$ loss function, also called $L_{0/1}$ loss in [23], which essentially counts the number of misclassified samples. More precisely, we take

$$\mathcal{L}_{0/1}(y, \tilde{f}(\mathbf{x})) := H(1 - y\tilde{f}(\mathbf{x})) \tag{5}$$

where H is the Heaviside unit step function

$$H(t) = \begin{cases} 1, & t > 0 \\ 0, & t \leq 0, \end{cases} \tag{6}$$

see the left panel of Figure 1.

However, it was also pointed out in [8] that the resulting optimization problem is *NP-complete, nonsmooth, and nonconvex* due to the discrete nature of the $(0,1)$ loss. The subsequent research focused on designing other simpler loss functions, notably convex ones like the *hinge* loss in which the step function $H(t)$ in (5) is replaced by t_+ in the right panel of Figure 1. However, the hinge loss has the drawback of being unbounded and thus is more sensitive to outliers in the data. Other convex and nonconvex loss functions for the SVM can be found in the excellent literature review in [23, Section 2].

In recent years, there has been a resurging interest in the original SVM problem with the $(0,1)$ loss, abbreviated as "$L_{0/1}$-SVM", following theoretical and algorithmic

Fig. 1: *Left*: the unit step function. *Right*: the function $t_+ := \max\{0, t\}$.

developments for optimization problems involving the "ℓ_0 *norm*", see e. g., [14, 24, 25, 27, 34, 35] and the references therein. In particular, the paper [23] proposed KKT-like optimality conditions for the *linear* $L_{0/1}$-SVM optimization problem and an efficient Alternating Direction Method of Multipliers (ADMM) solver to obtain an *approximate* solution whose performance is competitive among other SVM models. In this work, we draw inspiration from the aforementioned papers and present a *multiple-kernel* version of the theory in which the ambient functional space has a richer structure than the usual Euclidean space. More precisely, we shall formulate the $L_{0/1}$-SVM problem in the Multiple Kernel Learning (MKL) context, see e. g., [3, 12, 16, 20]. Clearly, the MKL framework can offer much more flexibility than the single-kernel formulation by letting the optimization algorithm determine the best combination of different kernel functions. In this sense, our results represent a substantial generalization of the work in [23] while maintaining the core features of the linear $L_{0/1}$-SVM.

The contributions of this paper are described next. We show that the optimal solutions of our MKL-$L_{0/1}$-SVM problem can also be characterized by a set of KKT-like conditions. The proofs are nontrivial extensions of those in [23] and rely on interplays between the infinite-dimensional and finite-dimensional problem formulations via the representer theorem [17]. Those optimality conditions can be readily exploited in designing an ADMM algorithm to solve the optimization problem despite the difficulties induced by the discontinuity and nonconvexity of the $(0, 1)$ loss function. Two working sets, one for the data and the other for the kernel combination, are employed to enhance the solution speed and to render sparsity in the solution. Convergence analysis of the ADMM is carried out under an additional assumption, and this point was completely missing in the linear case [23]. Moreover, any limit point generated by the algorithm is proved to be locally optimal. Finally, extensive numerical simulations on real data sets show that our MKL-$L_{0/1}$-SVM slightly outperforms the SimpleMKL approach in [16] in terms of accuracy of classification and the number of kernels used. Although a preliminary version of this paper [18] was presented at the 62nd IEEE Conference on Decision and Control (CDC 2023), here we lay the full theoretical foundation of the optimization problem (4), improve the algorithmic efficiency by the use of a finite-dimensional formulation, and provide much more numerical evidence for the competence of our approach.

The remainder of this paper is organized as follows. Section 2 reviews the classic $L_{0/1}$-SVM in the single-kernel case and discusses the MKL framework. Section 3 establishes the optimality theory for the MKL-$L_{0/1}$-SVM problem, and in Section 4 we propose an ADMM algorithm to solve the optimization problem. Section 5 provides convergence analysis of the ADMM. Finally, numerical experiments and concluding remarks are

provided in Sections 6 and 7, respectively.

Notation

The set of nonnegative reals is written as $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$, and $\mathbb{R}_+^n := \mathbb{R}_+ \times \cdots \times \mathbb{R}_+$ denotes the n-fold Cartesian product. We use $\mathbb{N}_m := \{1, 2, \ldots, m\}$ to represent the finite index set for the data points and $\mathbb{N}_L := \{1, 2, \ldots, L\}$ for the kernels where m and L are positive integers. Throughout the paper, the summation variable $i \in \mathbb{N}_m$ is reserved for the data index, and $\ell \in \mathbb{N}_L$ for the kernel index. We write $\sum_i$ and $\sum_\ell$ in place of $\sum_{i=1}^m$ and $\sum_{\ell=1}^L$ to simplify the notation.

2. PROBLEM FORMULATION

In all kernel-based methods, an important issue is to select a suitable kernel and its parameter which is also known as *hyperparameter*. Such a selection can be done via cross-validation once the functional form of the kernel is specified. In this context, Multiple Kernel Learning provides an alternative in which one employs a set of different kernels $\{\kappa_\ell\}_{\ell \in \mathbb{N}_L}$ and considers the SVM problem in the RKHS with a linearly combined kernel $\sum_\ell d_\ell \kappa_\ell$. Each $d_\ell \geq 0$ is a free parameter included in the optimization problem. In other words, one seeks to simultaneously find the best combination of the kernels and the optimal decision function via the solution of the MKL problem.

Before formally stating our MKL optimization problem for the $L_{0/1}$-SVM, we shall first briefly continue the discussion on the single-kernel problem in the Introduction, and then describe the necessary functional space setup borrowed from [16].

2.1. More on the single-kernel case

In this subsection, we further discuss the single-kernel version of the $L_{0/1}$-SVM along the lines of [18] and set up the notation. The optimization problem (4) is cast on the RKHS $\mathbb{H}$ which could be infinite-dimensional. However, one can appeal to the celebrated *representer theorem* [11] to reduce the problem to *finite* dimensions. More precisely, by the *semiparametric* representer theorem [17], any minimizer of (4) must have the form

$$\tilde{f}(\cdot) = \sum_i w_i \kappa(\,\cdot\,, \mathbf{x}_i) + b, \tag{7}$$

so that the desired function $w(\cdot)$ is completely parametrized by the m-dimensional vector $\mathbf{w} = \begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix}^\top$ which defines the linear combination of the *kernel sections* $\kappa(\,\cdot\,, \mathbf{x}_i)$. After some algebra involving the *kernel trick*, we obtain a finite-dimensional optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^m,\, b \in \mathbb{R}} \quad J(\mathbf{w}, b) := \frac{1}{2}\mathbf{w}^\top K \mathbf{w} + C \|(\mathbf{1} - A\mathbf{w} - b\mathbf{y})_+\|_0 \tag{8}$$

where,

- $K = K^\top$ is the *kernel matrix*

$$\begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_m, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix} \in \mathbb{R}^{m \times m} \tag{9}$$

  which is *positive semidefinite* by construction,

- $\mathbf{1} \in \mathbb{R}^m$ is a vector whose components are all 1's,

- $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}^\top$ is the vector of labels,

- the matrix $A = D_{\mathbf{y}} K$ is such that $D_{\mathbf{y}} = \operatorname{diag}(\mathbf{y})$ is the diagonal matrix whose (i, i) entry is y_i,

- the function $t_+ = \max\{0, t\}$ takes the positive part of the argument when applied to a scalar (right panel of Figure 1), and $\mathbf{v}_+ := \begin{bmatrix} (v_1)_+ & \cdots & (v_m)_+ \end{bmatrix}^\top$ represents componentwise application of the scalar function,

- $\|\mathbf{v}\|_0$ is the ℓ_0 norm[2] that counts the number of nonzero components in the vector $\mathbf{v}$. For a scalar $v \in \mathbb{R}$, we often write $|v|_0$ instead.

Clearly, the composite function $\|\mathbf{v}_+\|_0$ counts the number of positive components in $\mathbf{v}$. For a scalar t, the function $|t_+|_0$ coincides with the step function in (6).

Remark 2.1. The linear $L_{0/1}$-SVM studied in [23] can be viewed as a special case of the above problem where one employs a *homogeneous polynomial* kernel $\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y})^d$ with the degree parameter $d = 1$. In general, if the kernel $\kappa(\mathbf{x}, \mathbf{y})$ induces a *finite-dimensional* RKHS (as is the case for polynomial kernels), then the kernel matrix K, which is in fact a *Gram matrix*, must be *rank-deficient* when m is sufficiently large. In such a situation, we have the rank factorization $K = X^\top X$ such that $X \in \mathbb{R}^{r \times m}$ with $r = \operatorname{rank} K$. For example, in the previous case of $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$, we have $X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_m \end{bmatrix} \in \mathbb{R}^{n \times m}$ if such an X has full row rank ($r = n$). Then after a change of variables $\tilde{\mathbf{w}} = X\mathbf{w}$, the optimization problem (8) further reduces to

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^r,\, b \in \mathbb{R}} \quad \tilde{J}(\tilde{\mathbf{w}}, b) := \frac{1}{2}\|\tilde{\mathbf{w}}\|^2 + C\|(\mathbf{1} - \tilde{A}\tilde{\mathbf{w}} - b\mathbf{y})_+\|_0 \tag{10}$$

where $\tilde{A} = D_{\mathbf{y}} X^\top \in \mathbb{R}^{m \times r}$ is a tall and thin matrix. This is almost exactly the problem studied in [23].

For reasons discussed in Remark 2.1, in the remaining part of this paper, we shall always assume that the kernel matrix K is *positive definite*. This assumption holds in particular, for the *Gaussian* kernel

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), \tag{11}$$

[2]Since "ℓ_p norms" are not *bona fide* norms for $0 \le p < 1$, it may be better called ℓ_0 pseudonorm. However, we keep the terminology "ℓ_0 norm" for simplicity.

where $\sigma > 0$ is a hyperparameter, see [19]. In such a case, the matrix $A = D_{\mathbf{y}}K$ in (8) is also *invertible* since $D_{\mathbf{y}}$ is a diagonal matrix whose diagonal entries are the labels -1 or 1. Indeed, we have $D_{\mathbf{y}}^2 = D_{\mathbf{y}}^\top D_{\mathbf{y}} = I$.

2.2. Functional space for MKL

In order to enable MKL, we need to construct an RKHS whose kernel is a nonnegative linear combination of some given kernels $\{\kappa_\ell\}_{\ell\in\mathbb{N}_L}$. To this end, we shall follow the mathematical machinery in [16, Subsec. 2.1].

For each $\ell \in \mathbb{N}_L$, let $\mathbb{H}_\ell$ be an RKHS of functions on $\mathcal{X} \subset \mathbb{R}^n$ with the kernel $\kappa_\ell(\cdot,\cdot)$ and the inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}_\ell}$. Moreover, take $d_\ell \in \mathbb{R}_+$, and define a Hilbert space $\mathbb{H}'_\ell \subset \mathbb{H}_\ell$ as

$$\mathbb{H}'_\ell := \left\{ f \in \mathbb{H}_\ell : \frac{\|f\|_{\mathbb{H}_\ell}}{d_\ell} < \infty \right\} \tag{12}$$

endowed with the inner product

$$\langle f, g\rangle_{\mathbb{H}'_\ell} = \frac{\langle f, g\rangle_{\mathbb{H}_\ell}}{d_\ell}. \tag{13}$$

We use the convention that $x/0 = 0$ if $x = 0$ and ∞ otherwise. This means that, if $d_\ell = 0$ then a function $f \in \mathbb{H}_\ell$ belongs to the subspace $\mathbb{H}'_\ell$ only if $f = 0$. In such a case, $\mathbb{H}'_\ell$ becomes a trivial space containing only the null element. Within this framework, $\mathbb{H}'_\ell$ is an RKHS with the kernel $\kappa'_\ell(\mathbf{x},\mathbf{y}) = d_\ell\kappa_\ell(\mathbf{x},\mathbf{y})$ since

$$\forall f \in \mathbb{H}'_\ell \subset \mathbb{H}_\ell,\ f(\mathbf{x}) = \langle f(\cdot), \kappa_\ell(\mathbf{x},\cdot)\rangle_{\mathbb{H}_\ell} = \frac{1}{d_\ell}\langle f(\cdot), d_\ell\kappa_\ell(\mathbf{x},\cdot)\rangle_{\mathbb{H}_\ell} = \langle f(\cdot), d_\ell\kappa_\ell(\mathbf{x},\cdot)\rangle_{\mathbb{H}'_\ell}. \tag{14}$$

Define $\mathbb{F} := \mathbb{H}'_1 \times \mathbb{H}'_2 \times \cdots \times \mathbb{H}'_L$ as the Cartesian product of the RKHSs $\{\mathbb{H}'_\ell\}_{\ell\in\mathbb{N}_L}$, which is itself a Hilbert space with the inner product

$$\langle (f_1,\ldots,f_L), (g_1,\ldots,g_L)\rangle_{\mathbb{F}} = \sum_\ell \langle f_\ell, g_\ell\rangle_{\mathbb{H}'_\ell}. \tag{15}$$

Let $\mathbb{H} := \bigoplus_\ell \mathbb{H}'_\ell$ be the *direct sum* of the RKHSs $\{\mathbb{H}'_\ell\}_{\ell\in\mathbb{N}_L}$, which is also an RKHS with the kernel function

$$\kappa(\mathbf{x},\mathbf{y}) = \sum_\ell d_\ell\kappa_\ell(\mathbf{x},\mathbf{y}), \tag{16}$$

see [1]. Moreover, the squared norm of $f \in \mathbb{H}$ is known as

$$\|f\|_{\mathbb{H}}^2 = \min\left\{ \sum_\ell \|f_\ell\|_{\mathbb{H}'_\ell}^2 = \sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 : f = \sum_\ell f_\ell \text{ such that } f_\ell \in \mathbb{H}'_\ell \right\}. \tag{17}$$

The vector $\mathbf{d} = \begin{bmatrix} d_1 & \cdots & d_L \end{bmatrix}^\top \in \mathbb{R}_+^L$ is seen as a tunable parameter for the linear combination of kernels $\{\kappa_\ell\}_{\ell\in\mathbb{N}_L}$ in (16). It can be seen as an *adaptive* parameter that will enter the MKL optimization problem so that its value will be determined during the optimization process. Obviously, the precise linear combination of the kernels depends on the training data.

2.3. MKL-$L_{0/1}$-SVM optimization problem

We take inspiration from [16] and formulate the (soft-margin) MKL-$L_{0/1}$-SVM problem as

$$\begin{aligned}
\min_{\substack{\mathbf{f}=(f_1,\dots,f_L)\in\mathbb{F}\\ \mathbf{d}\in\mathbb{R}^L,\ b\in\mathbb{R}}} \quad & \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|^2_{\mathbb{H}_\ell} + C\sum_i \mathcal{L}_{0/1}(y_i, \tilde{f}(\mathbf{x}_i)) \\
\text{s.t.} \quad & d_\ell \geq 0,\ \ell \in \mathbb{N}_L && \text{(18a)}\\
& \sum_\ell d_\ell = 1 && \text{(18b)}\\
& \tilde{f}(\cdot) = \sum_\ell f_\ell(\cdot) + b
\end{aligned}$$

where $C > 0$ is a regularization parameter. The first (regularization) term in the objective function is chosen as such due to its convexity, see [16, Appendix A.1], which facilitates theoretical analysis. Moreover, constraints (18a) and (18b) define the standard $(L-1)$-simplex which ensures that the combined kernel is again positive definite. Also, the compactness property of the simplex is useful in the proof that a minimizer exists.

The last constraint in (18) can be safely eliminated by substituting it into the objective function. Next, let us define a new variable $\mathbf{u} \in \mathbb{R}^m$ by letting $u_i := 1 - y_i(f(\mathbf{x}_i) + b)$ where $f = \sum_\ell f_\ell$. Then the next lemma is a fairly straightforward result.

Lemma 2.2. The optimization problem (18) is equivalent to the following one:

$$\begin{aligned}
\min_{\substack{f\in\mathbb{H},\ \mathbf{d}\in\mathbb{R}^L\\ b\in\mathbb{R},\ \mathbf{u}\in\mathbb{R}^m}} \quad & \frac{1}{2}\|f\|^2_{\mathbb{H}} + C\|\mathbf{u}_+\|_0 && \text{(19a)}\\
\text{s.t.} \quad & d_\ell \geq 0,\ \ell \in \mathbb{N}_L \\
& \sum_\ell d_\ell = 1 \\
& u_i + y_i(f(\mathbf{x}_i) + b) = 1,\ i \in \mathbb{N}_m && \text{(19b)}
\end{aligned}$$

in the sense that they have the same set of minimizers once we introduce the same equality constraint for $\mathbf{u}$ in (18) and identify $f = \sum_\ell f_\ell$.

Proof. The proof relies on the relation (17). In view of that, we can rewrite the objective function of (19) as

$$\min\left\{\frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|^2_{\mathbb{H}_\ell} + C\|\mathbf{u}_+\|_0 : f = \sum_\ell f_\ell \text{ such that } f_\ell \in \mathbb{H}'_\ell\right\} \tag{20}$$

where the term $C\|\mathbf{u}_+\|_0$ can be seen as a constant with respect to this inner minimization.

Now suppose that the minimizer of (19) is $(f^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$. We will show that it is also a minimizer of (18). The converse can be handled similarly and is omitted. Let the

additive decomposition of f that achieves the minimum in (20) be $\sum_\ell \hat{f}_\ell$. Then, subject to the feasibility conditions, we have

$$\frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|\hat{f}_\ell^*\|_{\mathbb{H}_\ell}^2 + C\|\mathbf{u}_+^*\|_0 \le \frac{1}{2}\|f\|_{\mathbb{H}}^2 + C\|\mathbf{u}_+\|_0 \le \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 + C\|\mathbf{u}_+\|_0,$$

which means that $(\mathbf{f}^*, \mathbf{d}^*, b^*)$ is a minimizer of (18). □

3. OPTIMALITY THEORY

In this section, we give some theoretical results on the existence of an optimal solution to (19) and equivalently, to (18), and some KKT-like first-order optimality conditions. Our standing assumption is that each kernel matrix K_ℓ for the RKHS $\mathbb{H}_\ell$ is positive definite as e. g., in the case of Gaussian kernels with different hyperparameters. We state this formally below.

Assumption 1. Given the data points $\{\mathbf{x}_i : i \in \mathbb{N}_m\}$, each $m \times m$ kernel matrix K_ℓ, whose (i, j) entry is $\kappa_\ell(\mathbf{x}_i, \mathbf{x}_j)$, is positive definite for $\ell \in \mathbb{N}_L$.

We first provide an existence theorem for our MKL-$L_{0/1}$-SVM optimization problem whose proof can be found in the Appendix.

Theorem 3.1. Assume that the intercept b takes value from the closed interval $\mathcal{I} := [-M, M]$ where $M > 0$ is a sufficiently large number. Then the optimization problem (19) has a global minimizer and the set of all global minimizers is bounded.

The remaining part of this section is devoted to a characterization of global and local minimizers of the MKL-$L_{0/1}$-SVM problem. For the sake of consistency, let us rewrite (18) as follows:

$$\begin{aligned} \min_{\substack{\mathbf{f}\in\mathbb{F},\ \mathbf{d}\in\mathbb{R}^L \\ b\in\mathbb{R},\ \mathbf{u}\in\mathbb{R}^m}} \quad & \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 + C\|\mathbf{u}_+\|_0 && \text{(21a)} \\ \text{s.t.} \quad & d_\ell \ge 0,\ \ell \in \mathbb{N}_L \\ & \sum_\ell d_\ell = 1 \\ & u_i + y_i(f(\mathbf{x}_i) + b) = 1,\ i \in \mathbb{N}_m \end{aligned}$$

where the last equality constraint (identical to (19b)) is obviously *affine* in the "variables" $(\mathbf{f}, b, \mathbf{u})$, and we have identified $f = \sum_\ell f_\ell$ for simplicity. Before stating the optimality conditions, we need a generalized definition of a stationary point in nonlinear programming.

Definition 3.2. (P-stationary point of (21)**)** Fix a regularization parameter $C > 0$. We call $(\mathbf{f}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ a proximal stationary (abbreviated as P-stationary) point of the optimization problem (21) if there exists a vector $(\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^{L+1+m}$ and a number $\gamma > 0$ such that

$$d_\ell^* \ge 0,\ \ell \in \mathbb{N}_L \qquad \text{(22a)}$$

$$\sum_{\ell} d_\ell^* = 1 \tag{22b}$$

$$u_i^* + y_i \left(f^*(\mathbf{x}_i) + b^*\right) = 1,\ i \in \mathbb{N}_m \tag{22c}$$

$$\theta_\ell^* \geq 0,\ \ell \in \mathbb{N}_L \tag{22d}$$

$$\theta_\ell^* d_\ell^* = 0,\ \ell \in \mathbb{N}_L \tag{22e}$$

$$\forall \ell \in \mathbb{N}_L, \quad \frac{1}{d_\ell^*} f_\ell^*(\cdot) = -\sum_i \lambda_i^* y_i \kappa_\ell(\,\cdot\,, \mathbf{x}_i) \tag{22f}$$

$$-\frac{1}{2(d_\ell^*)^2} \|f_\ell^*\|_{\mathbb{H}_\ell}^2 + \alpha^* - \theta_\ell^* = 0,\ \ell \in \mathbb{N}_L \tag{22g}$$

$$\mathbf{y}^\top \boldsymbol{\lambda}^* = 0 \tag{22h}$$

$$\operatorname{prox}_{\gamma C \|(\cdot)_+\|_0}(\mathbf{u}^* - \gamma \boldsymbol{\lambda}^*) = \mathbf{u}^*, \tag{22i}$$

where the proximal operator is defined as

$$\operatorname{prox}_{\gamma C \|(\cdot)_+\|_0}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{v} \in \mathbb{R}^m} \quad C\|\mathbf{v}_+\|_0 + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{z}\|^2. \tag{23}$$

It can be shown [23] that the scalar version of the proximal operator above has a closed-form solution

$$\operatorname{prox}_{\gamma C |(\cdot)_+|_0}(z) := \operatorname*{argmin}_{v \in \mathbb{R}} \quad C|v_+|_0 + \frac{1}{2\gamma}(v - z)^2 = \begin{cases} 0, & 0 < z \leq \sqrt{2\gamma C} \\ z, & z > \sqrt{2\gamma C} \text{ or } z \leq 0, \end{cases} \tag{24}$$

see also Figure 2. The vector version of the proximal operator is evaluated by a componentwise application of (24), that is,

$$[\operatorname{prox}_{\gamma C \|(\cdot)_+\|_0}(\mathbf{z})]_i = \operatorname{prox}_{\gamma C |(\cdot)_+|_0}(z_i) \tag{25}$$

for $\mathbf{z} \in \mathbb{R}^m$ because the objective function on the right-hand side of (23) admits an additive componentwise decomposition. Formula (25) is called "$L_{0/1}$ proximal operator" in [23].

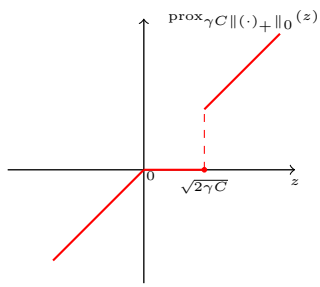

Fig. 2: The $L_{0/1}$ proximal operator on the real line.

The components of $(\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*)$ in Definition 3.2 can be understood as *Lagrange multipliers*, since they correspond to similar quantities in a smooth SVM problem [8]. However, here a dual problem seems difficult to derive because of the nonsmooth nonconvex

function $\|(\cdot)_+\|_0$. The set of equations and inequalities (22) are interpreted as *KKT-like* optimality conditions for the optimization problem (21), where (22a), (22b), and (22c) are the primal constraints, (22d) the dual constraints, (22e) the complementary slackness, and (22f), (22g), (22h), and (22i) the stationarity conditions of the Lagrangian with respect to the primal variables. We must point out that the nonsmoothness of the problem (21) is only found in the term $\|\mathbf{u}_+\|_0$, and the corresponding stationarity condition (22i) is given in terms of the proximal operator.

The following theorem connects the optimality conditions for (21) to P-stationary points, and the proof is deferred to the Appendix.

Theorem 3.3. The global and local minimizers of (21) admit the following characterizations:

(1) A global minimizer is a P-stationary point with $0 < \gamma < C_1$, where the positive number
$$C_1 := \min_{\mathbf{d}\in\mathbb{R}^L} \{\lambda_{\min}(\mathcal{K}(\mathbf{d})) : \mathbf{d} \text{ satisfies (18a) and (18b)}\} \tag{26}$$
in which $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix.

(2) Any P-stationary point (with $\gamma > 0$) is also a local minimizer of (21).

4. ALGORITHM DESIGN

In this section, we take advantages of the ADMM [6] and working sets (active sets) to devise a first-order algorithm for our MKL-$L_{0/1}$-SVM optimization problem. Before describing the algorithmic details, we must point out that the *finite-dimensional formulation* of the problem (see the proof of Theorem 3.1 in the Appendix) is more suitable for computation than the infinite-dimensional formulation (21). For this reason, we work on the former formulation in this section. The theory developed in Section 3, in particular, the notion of a P-stationary point, holds for the finite-dimensional problem modulo suitable adaptation which is done next.

First, let us rewrite the equivalent finite-dimensional MKL-$L_{0/1}$-SVM optimization problem as

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^m,\ \mathbf{d}\in\mathbb{R}^L \\ b\in\mathbb{R},\ \mathbf{u}\in\mathbb{R}^m}} \quad \frac{1}{2}\mathbf{w}^\top\mathcal{K}(\mathbf{d})\mathbf{w} + C\|\mathbf{u}_+\|_0 \tag{27a}$$

$$\text{s.t.} \quad d_\ell \geq 0,\ \ell \in \mathbb{N}_L$$

$$\sum_\ell d_\ell = 1$$

$$\mathbf{u} + \mathcal{A}(\mathbf{d})\mathbf{w} + b\mathbf{y} = \mathbf{1}. \tag{27b}$$

Definition 4.1. (P-stationary point of (27)**)** Fix a regularization parameter $C > 0$. We call $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ a P-stationary point of the optimization problem (27) if there exists a vector $(\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^{L+1+m}$ and a number $\gamma > 0$ such that the conditions (22a), (22b), (22d), (22e), (22h), (22i), and

$$\mathbf{u}^* + \mathcal{A}(\mathbf{d}^*)\mathbf{w}^* + b^*\mathbf{y} = \mathbf{1}, \tag{28a}$$

$$\mathbf{w}^* + D_{\mathbf{y}}\boldsymbol{\lambda}^* = \mathbf{0}, \tag{28b}$$

$$-\frac{1}{2}(\mathbf{w}^*)^\top K_\ell \mathbf{w}^* + \alpha^* - \theta_\ell^* = 0,\ \ell \in \mathbb{N}_L \tag{28c}$$

hold.

One can show that Definition 4.1 above and the previous Definition 3.2 are equivalent, as stated in the next proposition.

Proposition 4.2. Given a P-stationary point $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ in the sense of Definition 4.1, there corresponds a P-stationary point $(\mathbf{f}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ in the sense of Definition 3.2, and vice versa.

P r o o f . We shall first show that Definition 4.1 $\Longrightarrow$ Definition 3.2 . Given $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ and $(\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*)$ in Definition 4.1, we construct

$$f_\ell^*(\cdot) = d_\ell^* \sum_i w_i^* \kappa_\ell(\,\cdot\,, \mathbf{x}_i) \quad \forall \ell \in \mathbb{N}_L \tag{29}$$

which will be used to check the conditions in Definition 3.2. In view of the relation (28b), we immediately recover (22f). Moreover, from (29) we have

$$\|f_\ell^*\|_{\mathbb{H}_\ell}^2 = (d_\ell^*)^2 (\mathbf{w}^*)^\top K_\ell \mathbf{w}^*$$

which implies the equivalence between (22g) and (28c). The condition (22c) is equivalent to (28a) because the latter is simply a vectorized notation for the former. To see this point, just notice that the vector of function values $\begin{bmatrix} f^*(\mathbf{x}_1) & \cdots & f^*(\mathbf{x}_m) \end{bmatrix}^\top$ is equal to $\mathcal{K}(\mathbf{d}^*)\mathbf{w}^*$ in view of (29) and $f^* = \sum_\ell f_\ell^*$.

For the converse, i. e., Definition 3.2 $\Longrightarrow$ Definition 4.1, we simply define $\mathbf{w}^* := -D_{\mathbf{y}}\boldsymbol{\lambda}^*$, which is equivalent to (28b), given the vector $\boldsymbol{\lambda}^*$ in the sense of Definition 3.2. Then the above proof holds almost verbatim. □

Remark 4.3. The linear relation (28b) between $\boldsymbol{\lambda}^*$ and $\mathbf{w}^*$ is consistent with the proof of Theorem 3.3 (see Eq. (73) in the Appendix) and is analogous to the first formula in [23, Eq. (11)].

Now, we aim to obtain via numerical computation a P-stationary point in the sense of Definition 4.1 and hence a local minimizer of (27) by Theorem 3.3. To this end, let us first recognize a "working set" (with respect to the data) and its relation with *support vectors* following the development of [23, Subsec. 4.1], see also [18, Sec. V]. Suppose that $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ is a P-stationary point of (27). Then according to Definition 4.1, there exists a Lagrange multiplier $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ and a scalar $\gamma > 0$ such that (22i) holds. Define a set

$$T_* := \left\{ i \in \mathbb{N}_m : u_i^* - \gamma\lambda_i^* \in (0, \sqrt{2\gamma C}\,] \right\}, \tag{30}$$

and its complement $\overline{T}_* := \mathbb{N}_m \backslash T_*$. For a vector $\mathbf{z} \in \mathbb{R}^m$ and an index set $T \subset \mathbb{N}_m$ with cardinality $|T|$, we write $\mathbf{z}_T$ for the $|T|$-dimensional subvector of $\mathbf{z}$ whose components

are indexed in T. Then after some straightforward derivation using (22i), (24) and (25), we obtain $\mathbf{u}^*_{T_*} = \mathbf{0}$ and $\boldsymbol{\lambda}^*_{T_*} = \mathbf{0}$, and the working set (30) also admits the expression

$$T_* = \left\{ i \in \mathbb{N}_m : \lambda_i^* \in [-\sqrt{2C/\gamma}, 0) \right\}. \tag{31}$$

The formula tells us that the nonzero components of $\boldsymbol{\lambda}^*$ are indexed only in T_* with values in the interval $[-\sqrt{2C/\gamma}, 0)$. The following comments are immediate:

- The vectors $\{\mathbf{x}_i : i \in T_*\}$ corresponding to nonzero Lagrange multipliers $\{\lambda_i^*\}$ are called $L_{0/1}$-*support vectors* in [23]. They play the same role as standard support vectors in [8].
- Furthermore, the condition (22c) and $\mathbf{u}^*_{T_*} = \mathbf{0}$ imply that any $L_{0/1}$-support vector $\mathbf{x}_i$ satisfies one of the functional equations $f^*(\mathbf{x}) + b^* = \pm 1$. In the classic (linear) case, the two equations determine the *support hyperplanes*.

At this point, we are ready to give the framework of ADMM for the finite-dimensional optimization problem (27). In order to handle the inequality constraints (18a), we employ the indicator function (in the sense of Convex Analysis)

$$g(\mathbf{z}) = \begin{cases} 0, & \mathbf{z} \in \mathbb{R}^L_+ \\ +\infty, & \mathbf{z} \notin \mathbb{R}^L_+ \end{cases} \tag{32}$$

of the nonnegative orthant $\mathbb{R}^L_+$ and convert (27) to the form:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^m,\ \mathbf{d}\in\mathbb{R}^L \\ b\in\mathbb{R},\ \mathbf{u}\in\mathbb{R}^m,\ \mathbf{z}\in\mathbb{R}^L}} \quad \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d})\mathbf{w} + C\|\mathbf{u}_+\|_0 + g(\mathbf{z}) \tag{33a}$$

$$\text{s.t.} \quad \mathbf{d} - \mathbf{z} = \mathbf{0} \tag{33b}$$

$$\mathbf{1}^\top \mathbf{d} = 1 \tag{33c}$$

$$\mathbf{u} + \mathcal{A}(\mathbf{d})\mathbf{w} + b\mathbf{y} = \mathbf{1}, \tag{33d}$$

see [6, Section 5]. The *augmented Lagrangian* of (33) is given by

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{w}, \mathbf{d}, b, \mathbf{u}, \mathbf{z}; \boldsymbol{\theta}, \alpha, \boldsymbol{\lambda}) = {} & \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d})\mathbf{w} + C\|\mathbf{u}_+\|_0 + g(\mathbf{z}) + \boldsymbol{\theta}^\top(\mathbf{d} - \mathbf{z}) + \frac{\rho_2}{2}\|\mathbf{d} - \mathbf{z}\|^2 \\ & + \alpha\left(\mathbf{1}^\top \mathbf{d} - 1\right) + \frac{\rho_3}{2}\left(\mathbf{1}^\top \mathbf{d} - 1\right)^2 + \boldsymbol{\lambda}^\top\left[\mathbf{u} + \mathcal{A}(\mathbf{d})\mathbf{w} + b\mathbf{y} - \mathbf{1}\right] \\ & + \frac{\rho_1}{2}\|\mathbf{u} + \mathcal{A}(\mathbf{d})\mathbf{w} + b\mathbf{y} - \mathbf{1}\|^2 \end{aligned} \tag{34}$$

where $\boldsymbol{\theta}, \alpha, \boldsymbol{\lambda}$ are the Lagrange multipliers and each $\rho_j > 0$, $j = 1, 2, 3$, is a penalty parameter. Given the kth iterate $(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k)$, we use a Gauss–Seidel scheme to update each variable:

$$\mathbf{u}^{k+1} = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^m} \mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \tag{35a}$$

$$\mathbf{w}^{k+1} = \underset{\mathbf{w}\in\mathbb{R}^m}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{w}, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \tag{35b}$$

$$b^{k+1} = \underset{b\in\mathbb{R}}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \tag{35c}$$

$$\mathbf{z}^{k+1} = \underset{\mathbf{z}\in\mathbb{R}^L}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \tag{35d}$$

$$\mathbf{d}^{k+1} = \underset{\mathbf{d}\in\mathbb{R}^L}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \tag{35e}$$

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \rho_2(\mathbf{d}^{k+1} - \mathbf{z}^{k+1}) \tag{35f}$$

$$\alpha^{k+1} = \alpha^k + \rho_3\left(\mathbf{1}^\top \mathbf{d}^{k+1} - 1\right) \tag{35g}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho_1\left[\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^{k+1})\mathbf{w}^{k+1} + b^{k+1}\mathbf{y} - \mathbf{1}\right]. \tag{35h}$$

Next, we describe how to solve each subproblem above.

(1) **Updating $\mathbf{u}^{k+1}$.** The $\mathbf{u}$-subproblem in (35a) is equivalent to the following problem

$$\begin{aligned}\mathbf{u}^{k+1} &= \underset{\mathbf{u}\in\mathbb{R}^m}{\operatorname{argmin}}\ C\|\mathbf{u}_+\|_0 + \boldsymbol{\lambda}^\top\mathbf{u} + \frac{\rho_1}{2}\|\mathbf{u} + \mathcal{A}(\mathbf{d}^k)\mathbf{w}^k + b^k\mathbf{y} - \mathbf{1}\|^2 \\ &= \underset{\mathbf{u}\in\mathbb{R}^m}{\operatorname{argmin}}\ C\|\mathbf{u}_+\|_0 + \frac{\rho_1}{2}\|\mathbf{u} - \mathbf{s}^k\|^2 \\ &= \operatorname{prox}_{\frac{C}{\rho_1}\|(\cdot)_+\|_0}(\mathbf{s}^k),\end{aligned} \tag{36}$$

where

$$\mathbf{s}^k = \mathbf{1} - \mathcal{A}(\mathbf{d}^k)\mathbf{w}^k - b^k\mathbf{y} - \boldsymbol{\lambda}^k/\rho_1. \tag{37}$$

Define a working set T_k in the kth step by

$$T_k := \left\{i \in \mathbb{N}_m : s_i^k \in (0, \sqrt{2C/\rho_1})\right\}. \tag{38}$$

Then (36) can equivalently be written as

$$\mathbf{u}_{T_k}^{k+1} = \mathbf{0}, \quad \mathbf{u}_{\overline{T}_k}^{k+1} = \mathbf{s}_{\overline{T}_k}^k. \tag{39}$$

(2) **Updating $\mathbf{w}^{k+1}$.** The $\mathbf{w}$-subproblem in (35b) is

$$\begin{aligned}\mathbf{w}^{k+1} = \underset{\mathbf{w}\in\mathbb{R}^m}{\operatorname{argmin}}\ &\frac{1}{2}\mathbf{w}^\top\mathcal{K}(\mathbf{d}^k)\mathbf{w} + \boldsymbol{\lambda}^\top(\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w} + b^k\mathbf{y} - \mathbf{1}) \\ &+ \frac{\rho_1}{2}\|\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w} + b^k\mathbf{y} - \mathbf{1}\|^2.\end{aligned} \tag{40}$$

This is a quadratic minimization problem, and we only need to find a solution to the stationary-point equation

$$\mathcal{K}(\mathbf{d}^k)\mathbf{w} + \mathcal{A}(\mathbf{d}^k)^\top\boldsymbol{\lambda}^k + \rho_1\mathcal{A}(\mathbf{d}^k)^\top\left[\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w} + b^k\mathbf{y} - \mathbf{1}\right] = \mathbf{0} \tag{41}$$

which, after a rearrangement of terms, is a linear equation

$$\left[I + \rho_1\mathcal{K}(\mathbf{d}^k)\right]\mathbf{w} = -D_{\mathbf{y}}\left[\boldsymbol{\lambda}^k + \rho_1(\mathbf{u}^{k+1} + b^k\mathbf{y} - \mathbf{1})\right]. \tag{42}$$

The coefficient matrix is obviously positive definite and therefore a unique solution $\mathbf{w}^{k+1}$ exists.

(3) **Updating** b^{k+1}. The b-subproblem in (35c) can be written as

$$b^{k+1} = \underset{b\in\mathbb{R}}{\operatorname{argmin}} \ (\boldsymbol{\lambda}^k)^\top b\mathbf{y} + \frac{\rho_1}{2}\|\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w}^{k+1} + b\mathbf{y} - \mathbf{1}\|^2. \tag{43}$$

The stationary-point equation is

$$\mathbf{y}^\top\boldsymbol{\lambda}^k + \rho_1\mathbf{y}^\top(\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w}^{k+1} + b\mathbf{y} - \mathbf{1}) = 0, \tag{44}$$

and a solution is given by

$$b^{k+1} = -\frac{\mathbf{y}^\top\left[\boldsymbol{\lambda}^k + \rho_1(\mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^k)\mathbf{w}^{k+1} - \mathbf{1})\right]}{m\rho_1}. \tag{45}$$

(4) **Updating** $\mathbf{z}^{k+1}$. The subproblem for $\mathbf{z}$ in (35d) also admits a separation of variables, and we can carry out the update for each component as follows:

$$\begin{aligned} z_\ell^{k+1} &= \underset{z_\ell\in\mathbb{R}}{\operatorname{argmin}} \ \ g(\mathbf{z}) - (\boldsymbol{\theta}^k)^\top\mathbf{z} + \frac{\rho_2}{2}\|\mathbf{d}^k - \mathbf{z}\|^2 \\ &= \underset{z_\ell\in\mathbb{R}}{\operatorname{argmin}} \ \ g(z_\ell) - \theta_\ell^k z_\ell + \frac{\rho_2}{2}(d_\ell^k - z_\ell)^2 \\ &= (d_\ell^k + \theta_\ell^k/\rho_2)_+ \end{aligned} \tag{46}$$

where the function $(\cdot)_+$ takes the positive part of the argument. Inspired by this expression, we can define another working set

$$S_k := \left\{\ell \in \mathbb{N}_L : d_\ell^k + \theta_\ell^k/\rho_2 > 0\right\} \tag{47}$$

for the selection of kernels and $\overline{S}_k = \mathbb{N}_L\backslash S_k$. Then an equivalent update formula for $\mathbf{z}$ is

$$\mathbf{z}_{S_k}^{k+1} = (\mathbf{d}^k + \boldsymbol{\theta}^k/\rho_2)_{S_k}, \quad \mathbf{z}_{\overline{S}_k}^{k+1} = \mathbf{0}. \tag{48}$$

Notice that this working set is less complicated than T_k in (38) since the function $(\cdot)_+$ is continuous, unlike the proximal mapping (24).

(5) **Updating** $\mathbf{d}^{k+1}$. The $\mathbf{d}$-subproblem in (35e) can be written as

$$\begin{aligned} \mathbf{d}^{k+1} = \underset{\mathbf{d}\in\mathbb{R}^L}{\operatorname{argmin}} \ \ & \frac{1}{2}(\mathbf{w}^{k+1})^\top\sum_{\ell=1} d_\ell K_\ell\mathbf{w}^{k+1} + (\boldsymbol{\lambda}^k)^\top D_{\mathbf{y}}\sum_\ell d_\ell K_\ell\mathbf{w}^{k+1} \\ & + \frac{\rho_1}{2}\left\|\mathbf{u}^{k+1} + D_{\mathbf{y}}\sum_\ell d_\ell K_\ell\mathbf{w}^{k+1} + b^{k+1}\mathbf{y} - \mathbf{1}\right\|^2 \\ & + (\boldsymbol{\theta}^k)^\top\mathbf{d} + \frac{\rho_2}{2}\|\mathbf{d} - \mathbf{z}^{k+1}\|^2 + \alpha^k\mathbf{1}^\top\mathbf{d} + \frac{\rho_3}{2}\left(\mathbf{1}^\top\mathbf{d} - 1\right)^2. \end{aligned} \tag{49}$$

The stationary-point equation of the objective function with respect to each d_ℓ is

$$\begin{aligned} &\frac{1}{2}(\mathbf{w}^{k+1})^\top K_\ell\mathbf{w}^{k+1} + \rho_1(D_{\mathbf{y}}K_\ell\mathbf{w}^{k+1})^\top(\boldsymbol{\lambda}^k/\rho_1 + \mathbf{u}^{k+1} + b^{k+1}\mathbf{y} - \mathbf{1}) \\ &+ \rho_1(K_\ell\mathbf{w}^{k+1})^\top\sum_t d_t K_t\mathbf{w}^{k+1} + \theta_\ell^k + \rho_2(d_\ell - z_\ell^{k+1}) + \alpha^k + \rho_3\left(\mathbf{1}^\top\mathbf{d} - 1\right) = 0. \end{aligned} \tag{50}$$

We can rewrite it in matrix form as

$$\left(\rho_1 \mathfrak{A}^\top \mathfrak{A} + \rho_2 I + \rho_3 \mathbf{1}\cdot\mathbf{1}^\top\right)\mathbf{d} = \mathbf{v} - \boldsymbol{\theta}^k + \rho_2 \mathbf{z}^{k+1} + (\rho_3 - \alpha^k)\mathbf{1}, \tag{51}$$

where,

- $\mathfrak{A} = [K_1\mathbf{w}^{k+1}, \ldots, K_L\mathbf{w}^{k+1}] \in \mathbb{R}^{m\times L}$,
- and $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_L \end{bmatrix}^\top \in \mathbb{R}^L$ with

$$v_\ell = -\frac{1}{2}(\mathbf{w}^{k+1})^\top K_\ell \mathbf{w}^{k+1} - \rho_1 (D_{\mathbf{y}} K_\ell \mathbf{w}^{k+1})^\top (\boldsymbol{\lambda}^k/\rho_1 + \mathbf{u}^{k+1} + b^{k+1}\mathbf{y} - \mathbf{1}).$$

Clearly, the coefficient matrix in (51) is positive definite, and thus there exists a unique solution to the system of linear equations. However, it is possible that the solution vector could contain a negative component, which is not ideal because the components of $\mathbf{d}$ define the combination $\mathcal{K}(\mathbf{d})$ of the kernel matrices which must turn out to be positive definite. In order to handle such a pathology, we propose an ad-hoc recipe which projects the solution of the linear system to the standard $(L-1)$-simplex using the algorithm in [28]. Then motivated by the second formula in (48) and the equality constraint (33b), for $\ell \in \overline{S}_k$ we set

$$\mathbf{d}^{k+1}_{\overline{S}_k} = \mathbf{0}. \tag{52}$$

Remark 4.4. We have observed from our numerical simulations that the ad-hoc step of projection onto the simplex, which appears artificial, can significantly speed up convergence of the ADMM algorithm. In addition, the conference version of this paper [18] proposed a numerical procedure for the infinite-dimensional formulation (21), and the major difference is about the subproblems of $\mathbf{w}$ and $\mathbf{d}$. Indeed, each function f_ℓ in [18] is parametrized by its values at the data points $\{\mathbf{x}_i\}_{i\in\mathbb{N}_m}$, and the squared norm involves K_ℓ^{-1}. In this way, the vector of function values $\begin{bmatrix} f_\ell(\mathbf{x}_1) & \cdots & f_\ell(\mathbf{x}_m) \end{bmatrix}^\top$ has to be updated for each $\ell \in \mathbb{N}_L$. In contrast, the treatment in the current paper is more efficient and better conditioned since now only $\mathbf{w}$ needs to be updated thanks to the representer theorem and no explicit inversion is needed for the kernel matrices $\{K_\ell\}_{\ell\in\mathbb{N}_L}$. Moreover, the update of $\mathbf{d}$ is also greatly simplified to linear algebra, while previously one needed to solve a cubic polynomial equation for each component d_ℓ.

(6) **Updating $\boldsymbol{\theta}^{k+1}$.** With the help of the working set (47), the update of $\boldsymbol{\theta}$ in (35f) can be simplified as:

$$\boldsymbol{\theta}^{k+1}_{S_k} = \boldsymbol{\theta}^k_{S_k} + \rho_2(\mathbf{d}^{k+1} - \mathbf{z}^{k+1})_{S_k}, \quad \boldsymbol{\theta}^{k+1}_{\overline{S}_k} = \boldsymbol{\theta}^k_{\overline{S}_k}. \tag{53}$$

That is, the components of $\boldsymbol{\theta}$ outside the current working set S_k are not updated.

(7) **Updating α^{k+1}.** See (35g).

(8) **Updating $\boldsymbol{\lambda}^{k+1}$.** Inspired by the property of the working set T_* (see the line above (31)), the update of $\boldsymbol{\lambda}$ in (35h) is simplified as follows:

$$\boldsymbol{\lambda}_{T_k}^{k+1} = \boldsymbol{\lambda}_{T_k}^{k} + \rho_1 \mathbf{r}_{T_k}^{k+1}, \quad \boldsymbol{\lambda}_{\overline{T}_k}^{k+1} = \mathbf{0} \tag{54}$$

where the vector of residuals $\mathbf{r}^{k+1} := \mathbf{u}^{k+1} + \mathcal{A}(\mathbf{d}^{k+1})\mathbf{w}^{k+1} + b^{k+1}\mathbf{y} - \mathbf{1}$. In other words, we remove the components of $\boldsymbol{\lambda}$ which are not in the current working set.

The update steps above are summarized in Algorithm 1.

Algorithm 1 ADMM for the MKL-$L_{0/1}$-SVM

Require: C, ρ_1, ρ_2, ρ_3, $\{\kappa_\ell\}_{\ell\in\mathbb{N}_L}$, `max_iter`.

Set the counter $k = 0$, and initialize $(\mathbf{w}^0, \mathbf{d}^0, b^0, \mathbf{u}^0, \mathbf{z}^0; \boldsymbol{\theta}^0, \alpha^0, \boldsymbol{\lambda}^0)$.
while the terminating condition is not met and $k \le$ `max_iter` **do**
  Update T_k in (38) and $\mathbf{u}^{k+1}$ by (39).
  Update $\mathbf{w}^{k+1}$ by the unique solution to (42).
  Update b^{k+1} by (45).
  Update S_k in (47) and $\mathbf{z}^{k+1}$ by (48).
  Update $\mathbf{d}^{k+1}$ by the unique solution to (51) plus projection to the simplex and (52).
  Update $\boldsymbol{\theta}^{k+1}$ by (53).
  Update α^{k+1} by (35g).
  Update $\boldsymbol{\lambda}^{k+1}$ by (54).
  Set $k = k + 1$.
end while
return the final iterate $(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k)$.

Below we give a complexity analysis of the algorithm. In each round of the ADMM updates, we have the following observations:

- In the update of $\mathbf{u}_{k+1}$ by (39), the main term is $\mathcal{A}(\mathbf{d}^k)\mathbf{w}^k$ in (37) which takes $\mathcal{O}(Lm^2)$ operations.
- To update $\mathbf{w}^{k+1}$, the dominant computation is forming $\mathcal{K}(\mathbf{d}^k) = \sum_\ell d_\ell^k K_\ell$ and the solution of (42) whose time complexity is $\mathcal{O}(m^2 \max\{L, m\})$.
- Similarly, the product $\mathcal{A}(\mathbf{d}^k)\mathbf{w}^k$ is the most expensive computation in (45) to update b^{k+1}, and again its complexity is $\mathcal{O}(Lm^2)$.
- Updating $\mathbf{z}^{k+1}$ by (48) has a complexity $\mathcal{O}(L)$.
- To update $\mathbf{d}^{k+1}$, one first computes the columns of $\mathfrak{A}$ with $\mathcal{O}(Lm^2)$ operations. These columns can then be used to compute $\mathbf{v}$ with $\mathcal{O}(Lm)$ operations. The product $\mathfrak{A}^\top\mathfrak{A}$ needs L^2m operations, and the solution of the linear system takes $\mathcal{O}(L^3)$ operations. Therefore, the overall complexity is $\mathcal{O}(L \max\{m^2, Lm, L^2\}$.
- Computing $\boldsymbol{\theta}^{k+1}$ by (53) takes $\mathcal{O}(L)$ operations which is the same as $\mathbf{z}^{k+1}$, and updating α^{k+1} by (35g) takes $\mathcal{O}(1)$ operations.

- The complexity of updating $\boldsymbol{\lambda}^{k+1}$ is $\mathcal{O}(Lm^2)$, the same as b^{k+1}, due to the computation of the residual vector $\mathbf{r}^{k+1}$.

In summary, the time complexity for one round of updates of variables in Algorithm 1 is

$$\mathcal{O}(\max\{Lm^2, m^3, L^2m, L^3\}).$$

If the number of candidate kernels L is much smaller than m, then the above operations count reduces to $\mathcal{O}(m^3)$ which is reasonable.

5. CONVERGENCE ANALYSIS

In this section, we analyze certain convergence properties of Algorithm 1 proposed in the previous section. To ease the notation, let us write $\Psi^k = (\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k)$ for the kth iterate ($k \geq 1$) produced by the ADMM starting from some Ψ^0. The next assumption will be useful in this section.

Assumption 2. There exist a positive integer K and two sets $T_* \subset \mathbb{N}_m$ and $S_* \subset \mathbb{N}_L$ such that for all $k \geq K$, we have $T_k = T_*$ and $S_k = S_*$. That is, the working sets T_k in (38) and S_k in (47) remain unchanged after k becomes sufficiently large.

For simplicity, let us set $\rho_1 = \rho_2 = \rho_3 = \rho > 0$ in the augmented Lagrangian (34). We first establish a lemma that describes a *sufficient descent* property of the sequence $\{\mathcal{L}_\rho(\Psi^k)\}_{k\geq 1}$ of function values, and the proof is given in the Appendix.

Lemma 5.1. If the penalty parameter ρ is chosen sufficiently large, then there exist positive constants η and ε such that the inequality holds for all $k \geq 1$:

$$\begin{aligned}\mathcal{L}_\rho(\Psi^k) - \mathcal{L}_\rho(\Psi^{k+1}) \geq \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 \\ + \frac{\varepsilon}{2}\left[(\star^k) - \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2\right] + \frac{1}{\rho}(\blacktriangle^k),\end{aligned} \tag{55}$$

where $(\star^k) := \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 + m(b^{k+1} - b^k)^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \|\mathbf{d}^{k+1} - \mathbf{d}^k\|^2$, and $(\blacktriangle^k) := \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|^2 + (\alpha^{k+1} - \alpha^k)^2 + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2$.

In addition, suppose that Assumption 2 holds. Then we have a stronger inequality

$$\mathcal{L}_\rho(\Psi^k) - \mathcal{L}_\rho(\Psi^{k+1}) \geq \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \frac{\varepsilon}{2}(\star^k) + \frac{1}{\rho}(\blacktriangle^k) \tag{56}$$

which holds for all $k \geq K$.

Proposition 5.2. Choose the penalty parameter ρ sufficiently large as per Lemma 5.1, and suppose that Assumption 2 holds. In addition, assume that the sequence of dual variables $\{\boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k\}_{k\geq 1}$ is bounded. Then $\{\Psi^k\}_{k\geq 1}$ is also bounded.

Proof. Because of the boundedness assumption on the dual variables, the three quadratics in $\mathcal{L}_\rho$, namely $q_1(\mathbf{d}, \mathbf{z}) := \boldsymbol{\theta}^\top(\mathbf{d} - \mathbf{z}) + \frac{\rho_2}{2}\|\mathbf{d} - \mathbf{z}\|^2$, $q_2(\mathbf{d}) := \alpha\left(\mathbf{1}^\top\mathbf{d} - 1\right) +$

$\frac{\rho_3}{2}\left(\mathbf{1}^\top\mathbf{d}-1\right)^2$, and $q_3(\mathbf{w},\mathbf{d},b,\mathbf{u}) := \boldsymbol{\lambda}^\top[\mathbf{u}+\mathcal{A}(\mathbf{d})\mathbf{w}+b\mathbf{y}-\mathbf{1}]+\frac{\rho_1}{2}\|\mathbf{u}+\mathcal{A}(\mathbf{d})\mathbf{w}+b\mathbf{y}-\mathbf{1}\|^2$ are all bounded from below. Hence, by Lemma 5.1, we have

$$\infty > \mathcal{L}_\rho(\Psi_0) \geq \mathcal{L}_\rho(\Psi^k) \geq \frac{1}{2}(\mathbf{w}^k)^\top\mathcal{K}(\mathbf{d}^k)\mathbf{w}^k + C\|\mathbf{u}_+^k\|_0 + \text{const.} \tag{57}$$

In view of Assumption 1, letting $\|\mathbf{w}^k\| \to \infty$ will make $\mathcal{L}_\rho(\Psi^k) \to \infty$ which results in a contradiction. Therefore, $\{\mathbf{w}^k\}_{k\geq 1}$ remains bounded. The boundedness of b comes from the assumption in Theorem 3.1. The variable $\mathbf{d}$ is automatically bounded due to the simplex constraint. The boundedness of the remaining variables $\mathbf{z}$ and $\mathbf{u}$ can be shown through an argument similar to that after (57) with the quadratics q_1 and q_3 defined above. Consequently, all the primal variables are bounded, and this completes the proof. □

Appealing to the Bolzano–Weierstrass theorem, Proposition 5.2 implies that the sequence of ADMM iterates $\{\Psi^k\}_{k\geq 1}$ has a convergent subsequence $\{\Psi^{k_j}\}_{j\geq 1}$. The next result gives a weak form of convergence of $\{\Psi^k\}_{k\geq 1}$ in terms of the distance between two successive iterates.

Proposition 5.3. Under the assumptions of Proposition 5.2, it holds that

$$\lim_{k\to\infty}\|\Psi^{k+1}-\Psi^k\|^2 = 0, \tag{58}$$

where $\|\Psi\|^2$ is by definition equal to the sum of the squared norm of each variable in Ψ.

Proof. The proof of Proposition 5.2 shows that there exists a constant $C_2 \in \mathbb{R}$ such that $\mathcal{L}_\rho(\Psi) \geq C_2$. Moreover, by Lemma 5.1, the sequence of function values $\{\mathcal{L}_\rho(\Psi^k)\}_{k\geq 1}$ is monotonically decreasing. Therefore, $\mathcal{L}_\rho(\Psi^k)$ must converge to some real number as $k \to \infty$. The next chain of inequalities should be straightforward:

$$\begin{aligned}
\infty &> \mathcal{L}_\rho(\Psi^0) - C_2 \geq \mathcal{L}_\rho(\Psi^0) - \mathcal{L}_\rho(\Psi^N) \\
&= \sum_{k=0}^{N-1}\left[\mathcal{L}_\rho(\Psi^k) - \mathcal{L}_\rho(\Psi^{k+1})\right] \\
&\geq \sum_{k=0}^{N-1}\left[\frac{\eta}{2}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 + \frac{\varepsilon}{2}(\star^k) + \frac{1}{\rho}(\blacktriangle^k)\right] &\text{(59a)}\\
&= \frac{\eta}{2}\sum_{k=0}^{N-1}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 + \frac{\varepsilon}{2}\sum_{k=0}^{N-1}\left[\|\mathbf{u}^{k+1}-\mathbf{u}^k\|^2 + m(b^{k+1}-b^k)^2 + \|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2\right. \\
&\quad \left.+\|\mathbf{d}^{k+1}-\mathbf{d}^k\|^2\right] + \frac{1}{\rho}\sum_{k=0}^{N-1}\left[\|\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\|^2 + (\alpha^{k+1}-\alpha^k)^2 + \|\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k\|^2\right], &\text{(59b)}
\end{aligned}$$

where we have used Lemma 5.1 in (59a), and (59b) is just a substitution of $(\star^k)$ and $(\blacktriangle^k)$ defined in Lemma 5.1. Letting $N \to \infty$, we find that the series $\sum_{k=0}^{\infty}\|\mathbf{w}^{k+1}-$

$\mathbf{w}^k\|^2$, $\sum_{k=0}^{\infty}\|\mathbf{d}^{k+1}-\mathbf{d}^k\|^2$, $\sum_{k=0}^{\infty}(b^{k+1}-b^k)^2$, $\sum_{k=0}^{\infty}\|\mathbf{u}^{k+1}-\mathbf{u}^k\|^2$, $\sum_{k=0}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2$, $\sum_{k=0}^{\infty}\|\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\|^2$, $\sum_{k=0}^{\infty}(\alpha^{k+1}-\alpha^k)^2$, and $\sum_{k=0}^{\infty}\|\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k\|^2$ all have finite values, which yields the claim of this proposition. □

Remark 5.4. The full sequence convergence of Algorithm 1 seems very hard to prove in general due to the nonsmoothness and nonconvexity of the optimization problem (27). A line of research, see e. g., [2, 4, 5, 10, 13, 29, 30], for this type of problems imposes the *Kurdyka–Łojasiewicz (KŁ) property* for the objective function. However, due to the technical difficulty that in general, the sum of two KŁ functions is not guaranteed to be KŁ, it is not straightforward to show that our augmented Lagrangian (34) satisfies the KŁ property. For this reason, the issue of full sequence convergence will be left for future research.

The next result gives a characterization of the limit point *if* the algorithm converges, and the proof is left in the Appendix. We remark that in practice Algorithm 1 converges well as we have done extensive simulations to be presented in the next section.

Proposition 5.5. Suppose that the sequence $\{\Psi^k\}_{k\geq 1}$ generated by the ADMM algorithm above has a limit point $\Psi^* = (\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*, \mathbf{z}^*; \boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*)$. Then $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ is a P-stationary point in the sense of Definition 4.1 with $\gamma = 1/\rho_1$ and the Lagrange multipliers in the vector $(-\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*)$, and also a local minimizer of the problem (27).

Remark 5.6. The role of the working set S_k in (47) and its limit set S_* is to render *sparsity* in the combination of the kernels $\{\kappa_\ell\}_{\ell\in\mathbb{N}_L}$ for the MKL task, much similar to T_* in (30) and its effect on the support vectors. One can readily observe such sparsification from our simulation results in the next section, see the columns "# kernels" in Tables 2 and 3. This phenomenon can be explained by the following argument. Because of the nonnegativity condition (18a), the constraint (18b) can be interpreted as $\|\mathbf{d}\|_1 = 1$ where $\|\cdot\|_1$ denotes the ℓ_1 norm. The latter object is a convex relaxation of the ℓ_0 norm and is well known to be capable of inducing sparsity [9].

6. SIMULATIONS

In this section, we report results of numerical experiments using Matlab on a Dell laptop workstation with an Intel Core i7 CPU of 2.5 GHz on real data sets to demonstrate the effectiveness of the proposed MKL-$L_{0/1}$-SVM in comparison with the SimpleMKL approach of [16]. Our code can be found at `https://github.com/shiyj27/MKL-L01-SVM`.

(a) Stopping criterion. In our implementation, we terminate Algorithm 1 if the kth iterate $(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k)$ satisfies the condition

$$\max\left\{\beta_1^k, \beta_2^k, \beta_3^k, \beta_4^k, \beta_5^k, \beta_6^k, \beta_7^k, \beta_8^k\right\} < \texttt{tol}, \tag{60}$$

where the number $\texttt{tol} > 0$ is a tolerance level and

$$\begin{array}{llll} \beta_1^k := \|\mathbf{u}^k - \mathbf{u}^{k-1}\|, & \beta_2^k := \|\mathbf{w}^k - \mathbf{w}^{k-1}\|, & \beta_3^k := |b^k - b^{k-1}|, & \beta_4^k := \|\mathbf{z}^k - \mathbf{z}^{k-1}\| \\ \beta_5^k := \|\mathbf{d}^k - \mathbf{d}^{k-1}\|, & \beta_6^k := \|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\|, & \beta_7^k := |\alpha^k - \alpha^{k-1}|, & \beta_8^k := \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k-1}\|. \end{array} \tag{61}$$

This condition means, in plain words, that two successive iterates are sufficiently close.

(b) Parameters setting. In Algorithm 1, the parameters C and ρ_1 characterize the P-stationary-point condition (22i) (see Eq. (99) in the Appendix), and the working set (38) which is related to the number of support vectors. In order to choose the parameters C, ρ_1, ρ_2 and ρ_3, the standard 10-fold Cross Validation (CV) is employed on the training data, where the candidate values of C and ρ's are selected from the same set $\{2^{-2}, 2^{-1}, \ldots, 2^8\}$. The combination of parameters with the highest CV accuracy is picked out.

In addition, we choose $L = 10$ Gaussian kernels $\{\kappa_\ell\}_{\ell \in \mathbb{N}_L}$ with hyperparameters listed in Table 1 which are shared by both MKL-$L_{0/1}$-SVM and SimpleMKL for *all* data sets. We also set the maximum number of iterations $\texttt{max\_iter} = 10^3$ in Algorithm 1 and the tolerance level $\texttt{tol} = 10^{-3}$ in (60).

σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
0.1	0.2	0.3	0.5	0.7	1	1.2	1.5	1.7	2

Tab. 1: A quite arbitrary choice of the hyperparameters for 10 Gaussian kernels.

For the starting point, we set $\mathbf{w}^0 = \mathbf{u}^0 = \boldsymbol{\lambda}^0 = \mathbf{0}$, $\mathbf{z}^0 = \boldsymbol{\theta}^0 = \mathbf{0}$, $\alpha^0 = 0$ and $\mathbf{d}^0 = \frac{1}{L}\mathbf{1}$, $b^0 = 1$ or -1. The reason for such a choice is explained in the following. Let us refer to the objective function $J(\mathbf{w}, \mathbf{d}, b)$ in the finite-dimensional formulation of the MKL-$L_{0/1}$-SVM problem (see Eq. (64) in the Appendix). Then we immediately notice that $J(\mathbf{0}, \frac{1}{L}\mathbf{1}, 1) = Cm_-$ and $J(\mathbf{0}, \frac{1}{L}\mathbf{1}, -1) = Cm_+$ where m_+ and m_- denote the numbers of positive and negative components in the label vector $\mathbf{y}$. Therefore, we should choose $(\mathbf{w}^0, \mathbf{d}^0, b^0)$ such that $J(\mathbf{w}^0, \mathbf{d}^0, b^0) \leq C \min\{m_+, m_-\}$.

(c) Evaluation criterion. To evaluate the classification performance of our MKL-$L_{0/1}$-SVM, we use the testing accuracy ($\texttt{ACC}$):

$$\texttt{ACC} := 1 - \frac{1}{2m_{\text{test}}} \sum_{j=1}^{m_{\text{test}}} \left| \text{sign}\left(\sum_\ell f_\ell^*(\mathbf{x}^{\text{test}}) + b^* \right) - y_j^{\text{test}} \right|,$$

where $\{(\mathbf{x}_j^{\text{test}}, y_j^{\text{test}}) : j = 1, \ldots, m_{\text{test}}\}$ contains the testing data. Here, the quantity $\sum_\ell f_\ell^*(\mathbf{x}^{\text{test}})$ can be computed using (22f). More specifically, for each $\ell \in \mathbb{N}_L$ we can evaluate $f_\ell^*(\cdot) = -d_\ell^* \sum_i \lambda_i^* y_i \kappa_\ell(\,\cdot\,, \mathbf{x}_i)$ on the testing data using the convergent iterate produced by Algorithm 1.

(d) Simulation results. We have conducted experiments on four real data sets[3]. For each data set, standard feature normalization is performed. Then Algorithm 1 is run 30 times with different training and testing sets (70% of the data are randomly selected for training and the rest 30% for testing), and the last 20 $\texttt{ACC}$s are

[3]These data sets are downloaded from the UCI repository `https://archive.ics.uci.edu/`.

averaged. In order to avoid possibly bad local minimizers, we use a "warm start" strategy which initializes the algorithm with the convergent iterate of the previous run, hoping that the performance of classifier eventually becomes stable. For comparisons, we run the SimpleMKL algorithm[4] with the same ten Gaussian kernels (see Table 1 for the hyperparameters) under the same parameter configuration.

Sonar $m = 208$ $n = 60$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**1±0**	**76.5±3.7**	145±1.5	1.04±0.08
SimpleMKL	**1±0**	63.3±3.1	**145±0.5**	**0.05±0.02**

Wpbc $m = 198$ $n = 33$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**3.4±2.2**	**76.8±0.6**	138±0	0.65±0.15
SimpleMKL	8.9±0.73	76.7±0	**137±2.4**	**0.04±0.02**

Pima $m = 768$ $n = 8$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	2.4±0.8	73.3±4.1	490±34.3	106±58
SimpleMKL	**1±0**	**75.9±2.1**	**361±4.2**	**0.24±0.03**

Liver $m = 579$ $n = 9$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**1±0**	**71.7±0.8**	**405±0**	45±5.5
SimpleMKL	10±0.7	71.3±0	**405±0**	**0.43±0.05**

Tab. 2: Average performance measures for MKL-$L_{0/1}$-SVM and SimpleMKL on real data sets without additional outliers.

The simulation results are collected in Table 2 where one can see that our method achieves the best ACC on one data set "Sonar". While the ACCs on the data sets "Wpbc" and "Liver" are close for the two methods, our MKL-$L_{0/1}$-SVM requires much fewer kernels (see the column "# kernels") which significantly saves the computational cost for testing. In addition, the two methods perform similarly on the three data sets mentioned above in terms of the number of support vectors (NSV, sparsity in data usage) which is equal to $|T_*|$.

In order to see the influence of outliers in data sets on the two classifiers, we randomly flip the labels of 10% of the samples in each data set and redo the simulations. The updated results are reported in Table 3. It can be clearly seen that our MKL-$L_{0/1}$-SVM outperforms SimpleMKL in terms of the ACC on three data sets. On the fourth data set "Liver", the two methods have close ACCs, but our method still wins out in terms of the number of kernels on all four data sets. As for the computational time of training (solving the MKL optimization problems), we notice from the last columns of both Tables 2 and 3 that our method requires significantly more time because in general ADMM has *sublinear* convergence [6] while the gradient-type algorithm in [16] converges linearly.

[4]The code for SimpleMKL is found at https://github.com/shiyj27/SimpleMKL which is adapted for simulations reported in this paper from the source at https://github.com/maxis1718/SimpleMKL.

Sonar $m = 208$ $n = 60$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**1±0**	**74.29±2.2**	**144±0.5**	0.35±0.04
SimpleMKL	**1±0**	68.02±1.2	**144±0.5**	**0.05±0.02**

Wpbc $m = 198$ $n = 33$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**4.1±2.6**	**69.9±0.4**	**138±0**	0.70±0.34
SimpleMKL	6.6±2.8	69.8±0.5	138±1.2	**0.04±0.02**

Pima $m = 768$ $n = 8$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**3.4±1.1**	**69.3±2.7**	507±17.8	34.1±24.9
SimpleMKL	5.4±3.4	67.7±2.8	**413±23.8**	**0.31±0.05**

Liver $m = 579$ $n = 9$

Algorithm	# kernels	ACC (%)	NSV	time (s)
MKL-$L_{0/1}$-SVM	**1.8±0.4**	**66.8±1.1**	404±0.3	11.1±2.4
SimpleMKL	9.9±0.7	66.7±0	**404±0**	**0.47±0.07**

Tab. 3: Average performance measures for MKL-$L_{0/1}$-SVM and SimpleMKL on real data sets with 10% outliers by flipping the labels y_i.

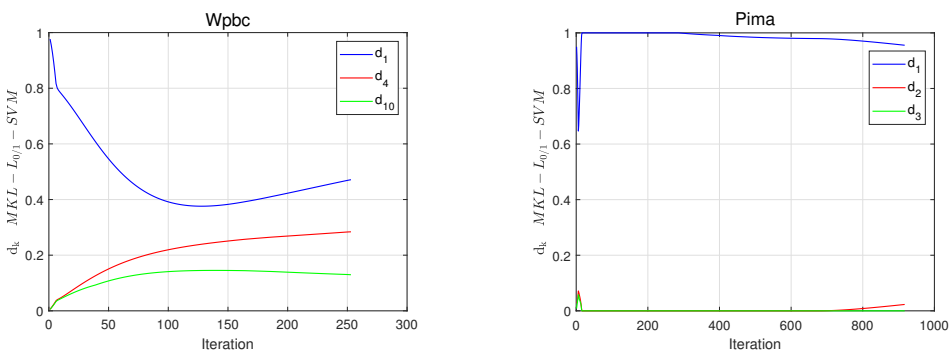

Fig. 3: Evolution of the three largest weights d_ℓ in MKL-$L_{0/1}$-SVM on two data sets. *Left*: Wpbc, *Right*: Pima.

Figure 3 depicts changes of $\mathbf{d}$ with respect to the number of ADMM iterations for the three largest components d_ℓ in one instance of Algorithm 1. For the "Pima" data set, in particular, the largest component d_1 is close to 1 when the algorithm converges.

7. CONCLUSION

In this paper, we have extended the $L_{0/1}$-SVM to a Multiple Kernel Learning (MKL) setting where the minimization of a regularized $(0, 1)$-loss function is accomplished together with the optimal combination of some given kernel functions. Despite the fact that the corresponding optimization problem is nonsmooth and nonconvex, we have

shown that the global and local minimizers can be characterized by a set of KKT-like optimality conditions. For the numerical solution of the MKL-$L_{0/1}$-SVM optimization problem, an efficient ADMM algorithm has been proposed to obtain a local optimum. Some convergence analysis for the ADMM is also provided. Simulation results on real data sets have manifested the effectiveness of our theory and algorithm.

Regarding future studies, second-order algorithms [26, 31, 32, 33, 34, 35] can be developed for the MKL-$L_{0/1}$-SVM optimization problem, since they can dramatically enhance the solution speed when the data size is not too large. In addition, it seems interesting to investigate the case of the problem in which Assumption 1 is violated, i. e., some kernel matrices are positive semidefinite and singular.

APPENDIX

Deferred Proofs

P r o o f. [Proof of Theorem 3.1] In view of Lemma 2.2, it is not difficult to argue that the optimization problem (18) is equivalent to

$$\begin{aligned} \min_{\mathbf{d}\in\mathbb{R}^L} \quad & \left\{ \min_{\substack{f\in\mathbb{H}\\ b\in\mathbb{R}}} \frac{1}{2}\|f\|_{\mathbb{H}}^2 + C\sum_i \| \left[1 - y_i(f(\mathbf{x}_i)+b)\right]_+ \|_0 \right\} \qquad (62a)\\ \text{s.t.} \quad & d_\ell \geq 0,\ \ell \in \mathbb{N}_L \\ & \sum_\ell d_\ell = 1. \end{aligned}$$

The inner minimization problem can be viewed as unconstrained once a feasible $\mathbf{d}$ is fixed. Hence we can employ the semiparametric representer theorem to conclude that the optimal f of the inner optimization has the form

$$f(\mathbf{x}) = \sum_\ell f_\ell(\mathbf{x}) = \sum_\ell d_\ell \sum_i w_i \kappa_\ell(\mathbf{x}, \mathbf{x}_i), \qquad (63)$$

where f_ℓ belongs to the RKHS $\mathbb{H}'_\ell$ with a kernel $d_\ell\kappa_\ell(\cdot,\cdot)$ for each $\ell \in \mathbb{N}_L$ (see Subsec. 2.2), and the parameter vector $\mathbf{w} = \begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix}^\top \in \mathbb{R}^m$. In view of this and (17), the optimization problem (62) is further equivalent to

$$\begin{aligned} \min_{\substack{\mathbf{w}\in\mathbb{R}^m,\ \mathbf{d}\in\mathbb{R}^L\\ b\in\mathbb{R}}} \quad & \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d})\mathbf{w} + C\|(\mathbf{1} - \mathcal{A}(\mathbf{d})\mathbf{w} - b\mathbf{y})_+\|_0 \qquad (64a)\\ \text{s.t.} \quad & d_\ell \geq 0,\ \ell \in \mathbb{N}_L \\ & \sum_\ell d_\ell = 1, \end{aligned}$$

where

$$\mathcal{K}(\mathbf{d}) := \sum_\ell d_\ell K_\ell \quad \text{and} \quad \mathcal{A}(\mathbf{d}) := D_{\mathbf{y}}\mathcal{K}(\mathbf{d}). \qquad (65)$$

Here K_ℓ is the kernel matrix corresponding to $\kappa_\ell(\cdot,\cdot)$. Let us write the objective function as

$$J(\mathbf{w},\mathbf{d},b) := \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d})\mathbf{w} + C\sum_i \left\|[(\mathbf{1}-\mathcal{A}(\mathbf{d})\mathbf{w}-b\mathbf{y})_i]_+\right\|_0. \tag{66}$$

It is obvious that the minimum value of $J(\mathbf{w},\mathbf{d},b)$ is no larger than $J(\mathbf{0},\mathbf{1}/L,0) = Cm$, where m is the maximum value of the ℓ_0 norm for vectors in $\mathbb{R}^m$. We can then consider the optimization problem (64) on the nonempty sublevel set

$$S := \{(\mathbf{w},\mathbf{d},b) \in \mathbb{R}^m \times \mathcal{I} \times \mathbb{R}^L : J(\mathbf{w},\mathbf{d},b) \le Cm\}. \tag{67}$$

We will prove that S is a compact set. Due to the finite dimensionality, it amounts to showing that S is closed and bounded. To this end, notice that the second term in the objective function (66) is lower-semicontinuous because so is the function $\|(\cdot)_+\|_0$ on $\mathbb{R}$ and the argument $(\mathbf{1}-\mathcal{A}(\mathbf{d})\mathbf{w}-b\mathbf{y})_i$ is smooth in $(\mathbf{w},\mathbf{d},b)$. We can now conclude that $J(\mathbf{w},\mathbf{d},b)$ is also lower-semicontinuous since the first quadratic term is smooth. Consequently, the sublevel set S is closed. The fact that S is also bounded follows from the argument that if we allow $\|\mathbf{w}\| \to \infty$, then

$$J(\mathbf{w},\mathbf{d},b) \ge \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d})\mathbf{w} \ge \frac{1}{2}\lambda_{\min}(\mathcal{K}(\mathbf{d}))\|\mathbf{w}\|^2 \to \infty, \tag{68}$$

where $\lambda_{\min}(\mathcal{K}(\mathbf{d})) > 0$ is the smallest eigenvalues of $\mathcal{K}(\mathbf{d})$. Therefore, the sublevel set S is compact and the existence of a global minimizer follows from the extreme value theorem of Weierstrass. The set of all global minimizers is a subset of S and is automatically bounded. □

Proof. [Proof of Theorem 3.3] To prove the first assertion, let $(\mathbf{f}^*,\mathbf{d}^*,b^*,\mathbf{u}^*)$ be a global minimizer of (21). If $\mathbf{u}^*$ is held fixed, then trivially $(\mathbf{f}^*,\mathbf{d}^*,b^*)$ must be a minimizer of the corresponding optimization problem with respect to the remaining variables $(\mathbf{f},\mathbf{d},b)$. In other words, we have

$$\begin{aligned}(\mathbf{f}^*,\mathbf{d}^*,b^*) = \underset{\substack{\mathbf{f}\in\mathbb{F},\ \mathbf{d}\in\mathbb{R}^L \\ b\in\mathbb{R}}}{\operatorname{argmin}} \quad & \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|^2_{\mathbb{H}_\ell} + C\|\mathbf{u}^*_+\|_0 \\ \text{s.t.} \quad & d_\ell \ge 0,\ \ell \in \mathbb{N}_L \\ & \sum_\ell d_\ell = 1 \\ & u_i^* + y_i\left(f(\mathbf{x}_i)+b\right) = 1 \quad \forall i.\end{aligned} \tag{69}$$

Since the term $C\|\mathbf{u}^*_+\|_0$ now becomes a fixed constant, the above problem turns into a *smooth convex* optimization problem, thanks to the result in [16, Appendix A.1]. Moreover, Slater's condition clearly holds, so we have *strong duality*. Consequently, conditions (22a) through (22h) hold as KKT conditions for (69).

It now remains to show the last condition (22i). This time we fix $\mathbf{d}^*$ and consider the minimization with respect to the other three variables. More precisely, we shall consider

the formulation derived from (64) with the additional variable $\mathbf{u}$:

$$\begin{aligned}(\mathbf{w}^*, b^*, \mathbf{u}^*) = \underset{\substack{\mathbf{w}\in\mathbb{R}^m,\ b\in\mathbb{R}\\ \mathbf{u}\in\mathbb{R}^m}}{\operatorname{argmin}} \quad & \frac{1}{2}\mathbf{w}^\top \mathcal{K}(\mathbf{d}^*)\mathbf{w} + C\|\mathbf{u}_+\|_0 \\ \text{s.t.} \quad & \mathbf{u} + \mathcal{A}(\mathbf{d}^*)\mathbf{w} + b\mathbf{y} = \mathbf{1}.\end{aligned} \tag{70}$$

Since the matrix $\mathcal{A}(\mathbf{d}^*) = D_{\mathbf{y}}\mathcal{K}(\mathbf{d}^*)$ is invertible by Assumption 1, we can write $\mathbf{w} = \mathcal{A}(\mathbf{d}^*)^{-1}(\mathbf{1} - \mathbf{u} - b\mathbf{y})$ and convert the optimization problem to another unconstrained version:

$$(\mathbf{u}^*, b^*) = \underset{\mathbf{u}\in\mathbb{R}^m,\, b\in\mathbb{R}}{\operatorname{argmin}} \quad g(\mathbf{u}, b) + C\|\mathbf{u}_+\|_0 \tag{71}$$

where $g(\mathbf{u}, b) := \frac{1}{2}(\mathbf{1} - \mathbf{u} - b\mathbf{y})^\top \mathcal{K}_1(\mathbf{d}^*)^{-1}(\mathbf{1} - \mathbf{u} - b\mathbf{y})$ is a new quadratic term such that the matrix $\mathcal{K}_1(\mathbf{d}) := D_{\mathbf{y}}\mathcal{K}(\mathbf{d})D_{\mathbf{y}}$. The following computation is straightforward:

$$\begin{aligned}\nabla g(\mathbf{u}, b) &= \begin{bmatrix}\nabla_{\mathbf{u}}\, g(\mathbf{u}, b) \\ \nabla_b\, g(\mathbf{u}, b)\end{bmatrix} = -\begin{bmatrix} I \\ \mathbf{y}^\top\end{bmatrix} \mathcal{K}_1(\mathbf{d}^*)^{-1}(\mathbf{1} - \mathbf{u} - b\mathbf{y}), \\ \nabla^2 g(\mathbf{u}, b) &= \begin{bmatrix} I \\ \mathbf{y}^\top\end{bmatrix} \mathcal{K}_1^{-1}(\mathbf{d}^*) \begin{bmatrix} I & \mathbf{y}\end{bmatrix}.\end{aligned} \tag{72}$$

Next, define

$$\boldsymbol{\lambda}^* := -\mathcal{K}_1(\mathbf{d}^*)^{-1}(\mathbf{1} - \mathbf{u}^* - b^*\mathbf{y}) = -\mathcal{K}_1(\mathbf{d}^*)^{-1}\mathcal{A}(\mathbf{d}^*)\mathbf{w}^* = -D_{\mathbf{y}}\mathbf{w}^*. \tag{73}$$

Take an arbitrary number $\gamma \in (0, C_1)$ and define a vector $\mathbf{z} := \operatorname{prox}_{\gamma C\|(\cdot)_+\|_0}(\mathbf{u}^* - \gamma\boldsymbol{\lambda}^*)$. We need to prove $\mathbf{z} = \mathbf{u}^*$ which is the last equation in (22). To this end, we shall emphasize three points:

(i) By the global optimality of $(\mathbf{u}^*, b^*)$, we have

$$g(\mathbf{u}^*, b^*) + C\|\mathbf{u}_+^*\|_0 \le g(\mathbf{z}, b^*) + C\|\mathbf{z}_+\|_0. \tag{74}$$

(ii) Using the second-order Taylor expansion for $g(\mathbf{u}, b)$, it is not difficult to show that

$$g(\mathbf{z}, b^*) - g(\mathbf{u}^*, b^*) \le (\boldsymbol{\lambda}^*)^\top(\mathbf{z} - \mathbf{u}^*) + \frac{\lambda_1}{2}\|\mathbf{z} - \mathbf{u}^*\|^2, \tag{75}$$

where $\lambda_1 := \lambda_{\max}(\mathcal{K}_1(\mathbf{d}^*)^{-1}) = 1/\lambda_{\min}(\mathcal{K}(\mathbf{d}^*))$.

(iii) By the definition of the proximal operator (23), we have

$$C\|\mathbf{z}_+\|_0 + \frac{1}{2\gamma}\|\mathbf{z} - (\mathbf{u}^* - \gamma\boldsymbol{\lambda}^*)\|^2 \le C\|\mathbf{u}_+^*\|_0 + \frac{1}{2\gamma}\|\gamma\boldsymbol{\lambda}^*\|^2 = C\|\mathbf{u}_+^*\|_0 + \frac{\gamma}{2}\|\boldsymbol{\lambda}^*\|^2. \tag{76}$$

Combining the above points via a chain of inequalities similar to [23, Eq. (S12), supplementary material], one can arrive at

$$0 \le \frac{\lambda_1 - 1/\gamma}{2}\|\mathbf{z} - \mathbf{u}^*\|^2 \le 0 \tag{77}$$

where the constant term in the middle is negative since we have chosen $0 < \gamma < C_1 \leq \lambda_{\min}(\mathcal{K}(\mathbf{d}^*) = 1/\lambda_1$. Therefore, we conclude that $\mathbf{z} = \mathbf{u}^*$, and we have completed the proof that a global minimizer $(\mathbf{f}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ of (21) is a P-stationary point.

To prove the second assertion, let us introduce the symbol $\phi = (\mathbf{f}, \mathbf{d}, b, \mathbf{u})$ for convenience. Suppose now that we have a P-stationary point $\phi^* = (\mathbf{f}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ with an associated vector $(\boldsymbol{\theta}^*, \alpha^*, \boldsymbol{\lambda}^*)$ and $\gamma > 0$. We need to show that ϕ^* is locally optimal in a sufficiently small neighborhood (to be constructed). To this end, let $U(\phi^*, \delta_1)$ be a sufficiently small neighborhood of ϕ^* of radius δ_1 such that for any $\phi = (\mathbf{f}, \mathbf{d}, b, \mathbf{u}) \in U(\phi^*, \delta_1)$ we have $u_i^* \neq 0 \implies u_i \neq 0$. That is to say: a small perturbation of $\mathbf{u}^*$ does not change the signs of its nonzero components. As a consequence, the relation

$$\|\mathbf{u}_+\|_0 \geq \|\mathbf{u}_+^*\|_0 \tag{78}$$

holds in that neighborhood. Next, notice that the first term $\frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2$ in the objective function (21a) is smooth in $(\mathbf{f}, \mathbf{d})$, which implies that it is locally Lipschitz continuous. Consequently, there exists a neighborhood $U(\phi^*, \delta_2)$ such that for any $\phi \in U(\phi^*, \delta_2)$, we have $\left|\frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 - \frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2\right| \leq C$ which further implies that

$$\frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2 - C \leq \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2. \tag{79}$$

Now, take $\delta = \min\{\delta_1, \delta_2\} > 0$ and consider the feasible region

$$\Theta := \{\phi = (\mathbf{f}, \mathbf{d}, b, \mathbf{u}) : (18a), (18b), \text{ and } (19b) \text{ hold}\} \tag{80}$$

of the problem (21). We will show that the P-stationary point ϕ^* is locally optimal in $\Theta \cap U(\phi^*, \delta)$, that is, $\phi \in \Theta \cap U(\phi^*, \delta)$ implies the inequality

$$\frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2 + C\left\|\mathbf{u}_+^*\right\|_0 \leq \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 + C\left\|\mathbf{u}_+\right\|_0. \tag{81}$$

For this purpose, let $\Gamma_* := \{i : u_i^* = 0\}$ and $\overline{\Gamma}_* := \mathbb{N}_m \backslash \Gamma_*$. Then by (22i) and the evaluation formulas for the proximal operator (see (24)and (25)), we have the following relation:

$$\begin{cases} i \in \Gamma_*,\ u_i^* = 0,\ -\sqrt{2C/\gamma} \leq \lambda_i^* < 0, \\ i \in \overline{\Gamma}_*,\ u_i^* \neq 0,\ \lambda_i^* = 0. \end{cases} \tag{82}$$

Consider the subset $\Theta_1 := \Theta \cap \{\phi : u_i \leq 0\ \forall i \in \Gamma_*\}$ of Θ, and the partition $\Theta = \Theta_1 \cup (\Theta \backslash \Theta_1)$. We will split the discussion into two cases given such a partition of the feasible region.

Case 1: $\phi \in \Theta_1 \cap U(\phi^*, \delta)$. We emphasize two conditions:

$$u_i \leq 0 \ \text{ for } \ i \in \Gamma_* \quad \text{and} \quad u_i + y_i\left(\sum_\ell f_\ell(\mathbf{x}_i) + b\right) = 1. \tag{83}$$

The following chain of inequalities holds:

$$\frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 - \frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2$$

$$\geq \frac{1}{2}\sum_{\ell}\left[\frac{1}{d_\ell^*}\langle 2f_\ell^*, f_\ell - f_\ell^*\rangle - \frac{d_\ell - d_\ell^*}{(d_\ell^*)^2}\|f_\ell^*\|_{\mathbb{H}_\ell}^2\right] \tag{84a}$$

$$= -\sum_i \lambda_i^* y_i\left[\sum_\ell f_\ell(\mathbf{x}_i) - \sum_\ell f_\ell^*(\mathbf{x}_i)\right] + \sum_\ell(\theta_\ell^* - \alpha^*)(d_\ell - d_\ell^*) \tag{84b}$$

$$= \sum_i \lambda_i^*(u_i - u_i^*) + \sum_\ell \theta_\ell^* d_\ell - \sum_\ell \theta_\ell^* d_\ell^* \tag{84c}$$

$$\geq \sum_i \lambda_i^*(u_i - u_i^*) = \sum_{i\in\Gamma_*}\lambda_i^* u_i \geq 0, \tag{84d}$$

where,

- (84a) is the first-order condition for convexity of each term $\frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2$,
- (84b) comes from the conditions (22f) and (22g) for the P-stationary point and the kernel trick,
- (84c) results from the feasibility condition on the right of (83), the fact that α^* is independent of the dummy variable ℓ, and the condition (18b),
- and (84d) is a consequence of (82) and the condition on the left of (83).

Combining (78) and (84), we obtain (81).

Case 2: $\phi \in (\Theta\backslash\Theta_1)\cap U(\phi^*,\delta)$. Note that $\phi\in\Theta\backslash\Theta_1$ means that there exists some $i\in\Gamma_*$ such that $u_i > 0$ while $u_i^* = 0$ (by the definition of Γ_*). Then we have $\|\mathbf{u}_+\|_0 \geq \|\mathbf{u}_+^*\|_0 + 1$. Combining this with (79), we have

$$\begin{aligned}\frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2 + C\left\|\mathbf{u}_+^*\right\|_0 \leq\ & \frac{1}{2}\sum_\ell \frac{1}{d_\ell^*}\|f_\ell^*\|_{\mathbb{H}_\ell}^2 + C\left\|\mathbf{u}_+\right\|_0 - C \\ \leq\ & \frac{1}{2}\sum_\ell \frac{1}{d_\ell}\|f_\ell\|_{\mathbb{H}_\ell}^2 + C\left\|\mathbf{u}_+\right\|_0,\end{aligned} \tag{85}$$

as desired. This completes the proof of local optimality of ϕ^*. □

Proof. [Proof of Lemma 5.1] As observed during the derivation of a solution to the $\mathbf{w}$-subproblem in (40), the augmented Lagrangian $\mathcal{L}_\rho$ in (34) is quadratic in the variable $\mathbf{w}$ when the rest variables are held fixed. In fact, the same observation can be made for each one of the other primal variables $\mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k$ if the nonsmooth part $C\|\mathbf{u}_+\|_0 + g(\mathbf{z})$ in $\mathcal{L}_\rho$ is not taken into consideration. Indeed, the nonsmooth part affects the updates of the variables $\mathbf{u}$ and $\mathbf{z}$ which must be handled with extra care, and this point will be addressed in the proof of the second part of the lemma about the stronger inequality (56).

Consider now the $\mathbf{u}$-update in (36). By definition, it must hold that

$$\mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \geq 0. \tag{86}$$

Next we analyze the quadratic minimization problem (40). One can compute the Hessian of $\mathcal{L}_\rho$ with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}}^2\mathcal{L}_\rho(\Psi) = \mathcal{K}(\mathbf{d}) + \rho\mathcal{A}(\mathbf{d})^\top\mathcal{A}(\mathbf{d}) = \mathcal{K}(\mathbf{d}) + \rho\mathcal{K}(\mathbf{d})^2, \tag{87}$$

where we have used (65) for the second equality. According to Assumption 1, we have $\nabla^2_{\mathbf{w}}\mathcal{L}_\rho(\Psi) \geq \eta I$ where $\eta := C_1 + \rho C_1^2 > 0$ with the positive constant C_1 defined in (26). Therefore, $\mathcal{L}_\rho$ is *η-strongly convex* [7] with respect to $\mathbf{w}$ which implies the inequality

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ \geq \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2.\end{aligned} \tag{88}$$

The analysis for the b-subproblem in (43) is similar. Indeed, the Hessian of $\mathcal{L}_\rho$ with respect to b reduces to the second-order derivative which is equal to $\rho\mathbf{y}^\top\mathbf{y} = m\rho$. Now, the constant Hessian leads to the equality

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ = \frac{m\rho}{2}(b^{k+1} - b^k)^2.\end{aligned} \tag{89}$$

The $\mathbf{z}$-update in (46) by definition implies that

$$\mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \geq 0. \tag{90}$$

For the $\mathbf{d}$-subproblem in (49) which is again quadratic, we can also compute the Hessian $\nabla^2_{\mathbf{d}}\mathcal{L}_\rho(\Psi)$ which is equal to the coefficient matrix on the left-hand side of (51). Hence $\nabla^2_{\mathbf{d}}\mathcal{L}_\rho(\Psi) \geq \rho I$ which gives rise to

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^{k+1}, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ \geq \frac{\rho}{2}\|\mathbf{d}^{k+1} - \mathbf{d}^k\|^2.\end{aligned} \tag{91}$$

Since $\mathcal{L}_\rho$ depends on the dual variables $\boldsymbol{\theta}, \alpha, \boldsymbol{\lambda}$ linearly, the updates of these variables in (35f), (35g), and (35h) can be combined to yield the equality

$$\begin{aligned}&\mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^{k+1}, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ &\qquad - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^{k+1}, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^{k+1}, \alpha^{k+1}, \boldsymbol{\lambda}^{k+1}) = -\frac{1}{\rho}(\blacktriangle^k).\end{aligned} \tag{92}$$

Adding up the formulas (86), (88), (89), (90), (91), and (92), we obtain

$$\begin{aligned}\mathcal{L}_\rho(\Psi^k) - \mathcal{L}_\rho(\Psi^{k+1}) &\geq \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \frac{\rho}{2}(\underline{\star}^k) - \frac{1}{\rho}(\blacktriangle^k) \\ &= \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \frac{\varepsilon}{2}(\underline{\star}^k) + \frac{1}{\rho}(\blacktriangle^k) + \frac{\rho - \varepsilon}{2}(\underline{\star}^k) - \frac{2}{\rho}(\blacktriangle^k)\end{aligned} \tag{93}$$

where $(\underline{\star})^k := m(b^{k+1} - b^k)^2 + \|\mathbf{d}^{k+1} - \mathbf{d}^k\|^2 = (\star^k) - \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2$ and $\varepsilon > 0$ is a sufficient small number relative to ρ. Now we impose the condition

$$\rho > \sup_{k\geq 1} \frac{2(\blacktriangle^k)^{1/2}}{(\underline{\star}^k)^{1/2}}. \tag{94}$$

Because of the strict inequality, there exists $\varepsilon > 0$ such that the left-hand side of (94) can be replaced with $\rho - \varepsilon$. A consequence of these two inequalities is that the last two terms in (93) satisfy $\frac{\rho-\varepsilon}{2}(\bigstar^k) - \frac{2}{\rho}(\blacktriangle^k) \geq 0$ for all $k \geq 1$ which implies the weak inequality (55).

It remains to show the strong inequality (56) under Assumption 2. In view of the additional assumption on the index sets T_k and the $L_{0/1}$ proximal operator (24), the **u**-update in (36) implies that for $k \geq K$,

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^k, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^k, \mathbf{d}^k, b^k, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ = \frac{\rho}{2} \sum_{i \in \overline{T}_*} (u_i^{k+1} - u_i^k)^2 = \frac{\rho}{2} \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2, \end{aligned} \tag{95}$$

where the second equality follows from the fact that $\mathbf{u}_{T_*}^{k+1} = \mathbf{u}_{T_*}^k = \mathbf{0}$, see (39). In a similar fashion, the **z**-update in (46) leads to equality

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^k; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\rho(\mathbf{w}^{k+1}, \mathbf{d}^k, b^{k+1}, \mathbf{u}^{k+1}, \mathbf{z}^{k+1}; \boldsymbol{\theta}^k, \alpha^k, \boldsymbol{\lambda}^k) \\ = \frac{\rho}{2} \sum_{\ell \in S_*} (z_\ell^{k+1} - z_\ell^k)^2 = \frac{\rho}{2} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \end{aligned} \tag{96}$$

where the part of Assumption 2 on S_k plays a role. Now the desired inequality (56) can be established using an argument similar to the one leading to (55). □

P r o o f. [Proof of Proposition 5.5] Let $\{\Psi^j \equiv \Psi^{k_j}\}_{j \geq 1}$ be a subsequence of $\{\Psi^k\}_{k \geq 1}$ that converges to Ψ^*, and let $\mathbf{s}^*$ be the limit of the corresponding subsequence of $\{\mathbf{s}^k\}_{k \geq 1}$ in (37). The simplest case is the dual update for α^k. Since the constants ρ_j, $j = 1, 2, 3$ are positive, we have $\mathbf{1}^\top \mathbf{d}^* = 1$, which is the condition (22b), by taking the limit of the α update (35g).

Now look at the limit of the dual update for $\boldsymbol{\lambda}$. Define T_* to be the limit working set in the sense of (38) with $\mathbf{s}^k$ replaced by $\mathbf{s}^*$. After taking the limit of the update formula (54), we obtain

$$\mathbf{r}_{T_*}^* = \mathbf{0} \quad \text{and} \quad \boldsymbol{\lambda}_{\overline{T}_*}^* = \mathbf{0}. \tag{97}$$

In addition, we check the limit of (39) and get $\mathbf{u}_{T_*}^* = \mathbf{0}$ and $\mathbf{u}_{\overline{T}_*}^* = \mathbf{s}_{\overline{T}_*}^* = (\mathbf{u}^* - \mathbf{r}^* - \boldsymbol{\lambda}^*/\rho_1)_{\overline{T}_*}$ where the last equality comes from the relation

$$\mathbf{s}^k = \mathbf{u}^k - \mathbf{r}^k - \boldsymbol{\lambda}^k/\rho_1. \tag{98}$$

With reference to (97), plain computation yields $\mathbf{r}_{\overline{T}_*}^* = \mathbf{0}$ and furthermore $\mathbf{r}^* = \mathbf{0}$. This is the condition (28a). Now (98) reduces to $\mathbf{s}^* = \mathbf{u}^* - \boldsymbol{\lambda}^*/\rho_1$ in the limit. It is then not difficult to check that the update formula (36) also holds in the limit which can be written as

$$\mathbf{u}^* = \operatorname{prox}_{\frac{C}{\rho_1}\|(\cdot)_+\|_0}(\mathbf{u}^* - \boldsymbol{\lambda}^*/\rho_1). \tag{99}$$

This is the condition (22i) with $\gamma = 1/\rho_1$, and also explains why here we have used the same symbol T_* as in (30).

Next we take the limit of (41). Since we have shown (28a), the last term in (41) vanishes in the limit and we are left with (28b). Similarly, taking the limit of (44) leads

to (22h). In order to describe the limit of (48), let us introduce the limit working set S_* in the sense of (47) with d_ℓ^k and θ_ℓ^k replaced by d_ℓ^* and θ_ℓ^*, respectively. Then we have

$$\mathbf{z}^*_{S_*} = (\mathbf{d}^* + \boldsymbol{\theta}^*/\rho_2)_{S_*}, \quad \mathbf{z}^*_{\overline{S}_*} = \mathbf{0}. \tag{100}$$

Moreover, taking the limit of $\boldsymbol{\theta}^k$ in (53) gives $\mathbf{d}^*_{S_*} = \mathbf{z}^*_{S_*}$ which is a part of the constraint (33b), and together with (100) implies $\boldsymbol{\theta}^*_{S_*} = \mathbf{0}$. In addition, the last update formula (52) for $\mathbf{d}$ implies $\mathbf{d}^*_{\overline{S}_*} = \mathbf{0}$. These two relations establish the complementary slackness condition (22e) and the desired equality $\mathbf{d}^* = \mathbf{z}^*$. Also, the update strategy for $\mathbf{d}$ ensures that each d_ℓ^k is nonnegative which yields the limit (22a). In order to obtain (28c), we take the limit of the stationary-point equation (50), notice that the terms involving ρ_1, ρ_2, and ρ_3 vanish due to the established optimality conditions, and take (28b) into account. Here we have a sign difference for the Lagrange multiplier $\boldsymbol{\theta}$.

The only missing condition up till now is (22d) in the index set $\overline{S}_*$. By definition, for $\ell \in \overline{S}_*$ it holds that $d_\ell^* + \theta_\ell^*/\rho_2 \le 0$ which implies $\theta_\ell^* \le -\rho_2 d_\ell^* \le 0$. This is precisely (22d) with a sign difference. Therefore, $(\mathbf{w}^*, \mathbf{d}^*, b^*, \mathbf{u}^*)$ is a P-stationary point with $\gamma = 1/\rho_1$ and its local optimality is guaranteed by Theorem 3.3. □

ACKNOWLEDGMENT

This work was partially supported by the Shenzhen Science and Technology Program (Grant No. 202206193000001-20220817184157001), the Fundamental Research Funds for the Central Universities, and the "Hundred-Talent Program" of Sun Yat-sen University.

The authors would also like to thank Mr. Jiahao Liu and Mr. Ji Cheng for their assistance in the Matlab implementation of Algorithm 1.

(Received June 27, 2025)

REFERENCES

[1] N. Aronszajn: Theory of reproducing kernels. Trans. American Math. Soc. *68* (1950), 337–404. DOI:10.1090/S0002-9947-1950-0051437-7

[2] H. Attouch, J. Bolte, and B. F. Svaiter: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Programming *137* (2013), 91–129. DOI:10.3917/ris.091.0129

[3] F. R. Bach, G. R. Lanckriet, and M. I. Jordan: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proc. 21st International Conference on Machine Learning, 2004, pp. 41–48.

[4] J. Bolte, S. Sabach, and M. Teboulle: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Programming *146* (2014), 59–494. DOI:10.1007/s10107-013-0701-9

[5] R. I. Boţ and D. K. Nguyen: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. *45* (2020), 682–712. DOI:10.1287/moor.2019.1008

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning *3* (2011), 1–122.

[7] S. P. Boyd and L. Vandenberghe: Convex Optimization. Cambridge University Press, 2004. DOI:10.1017/S275390670000070X

[8] C. Cortes and V. Vapnik: Support-vector networks. Machine Learning *20* (1995), 273–297. DOI:10.1023/A:1022627411411

[9] T. Hastie, R. Tibshirani, and M. Wainwright: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton 2015.

[10] M. Hong, Z. Q. Luo, and M. Razaviyayn: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. *26* (2016), 337–364. DOI:10.1137/140990309

[11] G. Kimeldorf and G. Wahba: Some results on Tchebycheffian spline functions. J. Math. Analysis Appl. *33* (1971), 82–95. DOI:10.1016/0022-247X(71)90184-3

[12] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan: Learning the kernel matrix with semidefinite programming. J. Machine Learn. Res. *5* (2004), 27–72.

[13] G. Li and T. K. Pong: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. *25* (2015), 2434–2460. DOI:10.1137/140998135

[14] M. Nikolova: Description of the minimizers of least squares regularized with ℓ_0-norm. uniqueness of the global minimizer. SIAM J. Imaging Sci. *6* (2013), 904–937. DOI:10.1137/11085476X

[15] V. I. Paulsen and M. Raghupathi: An Introduction to the Theory of Reproducing Kernel Hilbert Spaces. Cambridge Studies in Advanced Mathematics *152*, Cambridge University Press, Cambridge 2016.

[16] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet: SimpleMKL. J. Machine Learn. Res. *9* (2008), 2491–2521.

[17] B. Schölkopf and A. J. Smola: Learning with Kernels. Adaptive Computation and Machine Learning *4*, MIT Press, Cambridge 2001.

[18] Y. Shi and B. Zhu: An ADMM solver for the MKL-$L_{0/1}$-SVM. In: Proc. 62nd IEEE Conference on Decision and Control (CDC 2023), Singapore 2023, pp. 3339–3346.

[19] K. Slavakis, P. Bouboulis, and S. Theodoridis: Online learning in reproducing kernel Hilbert spaces. In: Academic Press Library in Signal Processing *1*, Academic Press, Cambridge 2014, pp. 883–987.

[20] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf: Large scale multiple kernel learning. J. Machine Learn. Res. *7* (2006), 1531–1565.

[21] S. Theodoridis: Machine Learning: A Bayesian and Optimization Perspective. (Second edition.) Academic Press, Cambridge 2020.

[22] V. Vapnik: The Nature of Statistical Learning Theory. (Second edition.) Springer Science and Business Media, New York 2000.

[23] H. Wang, Y. Shao, S. Zhou, C. Zhang, and N. Xiu: Support vector machine classifier via $L_{0/1}$ soft-margin loss. IEEE Trans. Pattern Anal. Machine Intell.*44* (2022), 7253–7265. DOI:10.1109/TPAMI.2021.3092177

[24] L. Wang, W. Liu, and B. Zhu: ADMM for ℓ_0 factor analysis. In: Proc. 13th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2024), Corvallis 2024.

[25] L. Wang, W. Liu, and B. Zhu: ℓ_0 factor analysis. In: Proc. 63rd IEEE Conference on Decision and Control (CDC 2024), Milan 2024, pp. 7214–7219.

[26] L. Wang, W. Liu, and B. Zhu: A Newton interior-point method for ℓ_0 factor analysis. In: Proc. 64th IEEE Conference on Decision and Control (CDC 2025), Rio de Janeiro 2025, pp. 3232–3237.

[27] L. Wang, B. Zhu, and W. Liu: ℓ_0 factor analysis: A P-stationary point theory. IEEE Trans. Automat. Control *70* (2025), 6050–6063. DOI:10.1109/TAC.2025.3552869

[28] W. Wang and M. A. Carreira-Perpinán: Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. arXiv preprint: 1309.1541, 2013.

[29] Y. Wang, W. Yin, and J. Zeng: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Scientific Computing *78* (2019), 29–63. DOI:10.1007/s10915-018-0757-z

[30] L. Yang, T. K. Pong, and X. Chen: Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. SIAM J. Imaging Sci. *10* (2017), 74–110. DOI:10.1137/15M1027528

[31] Y. Zhang, N. Zhang, D. Sun, and K. C. Toh: A proximal point dual Newton algorithm for solving group graphical Lasso problems. SIAM J. Optim. *30* (2020), 2197–2220. DOI:10.1137/19M1267830

[32] S. Zhou: Gradient projection Newton pursuit for sparsity constrained optimization. Appl. Comput. Harmonic Anal. *61* (2022), 75–100. DOI:10.1016/j.acha.2022.06.002

[33] S. Zhou, L. Pan, and N. Xiu: Newton method for ℓ_0-regularized optimization. Numer. Algorithms (2021), 1–30.

[34] S. Zhou, L. Pan, N. Xiu, and H. D. Qi: Quadratic convergence of smoothing Newton's method for 0/1 loss optimization. SIAM J. Optim. *31* (2021), 3184–3211. DOI:10.1137/21M1409445

[35] S. Zhou, N. Xiu, and H. D. Qi: Global and quadratic convergence of Newton hard-thresholding pursuit. J. Machine Learn. Res. *22* (2021), 1–45.

Bin Zhu, School of Intelligent Systems Engineering, Sun Yat-sen University, Gongchang Road 66, Shenzhen 518107. P. R. China.
e-mail: zhub26@mail.sysu.edu.cn

Yijie Shi, Xi'an XD High Voltage Switchgear Operating Mechanism Co., Ltd., Dengling Road 24, Xi'an 710076. P. R. China.
e-mail: 459258149@qq.com