Matrix-Regularized Multiple Kernel Learning via (r, p) Norms

Yina Han, *Member, IEEE*, Yixin Yang, *Member, IEEE*, Xuelong Li, *Fellow, IEEE*,
Qingyu Liu, and Yuanliang Ma

***Abstract*— This paper examines a matrix-regularized multiple kernel learning (MKL) technique based on a notion of (r, p) norms. For the problem of learning a linear combination in the support vector machine-based framework, model complexity is typically controlled using various regularization strategies on the combined kernel weights. Recent research has developed a generalized ℓ_p-norm MKL framework with tunable variable $p (p \geq 1)$ to support controlled intrinsic sparsity. Unfortunately, this "1-D" vector ℓ_p-norm hardly exploits potentially useful information on how the base kernels "interact." To allow for higher order kernel-pair relationships, we extend the "1-D" vector ℓ_p-MKL to the "2-D" matrix (r, p) norms ($1 \leq r, p < \infty$). We develop a new formulation and an efficient optimization strategy for (r, p)-MKL with guaranteed convergence. A theoretical analysis and experiments on seven UCI data sets shed light on the superiority of (r, p)-MKL over ℓ_p-MKL in various scenarios.**

***Index Terms*— Generalization bound, matrix regularization, multiple kernel learning (MKL), support vector machine (SVM).**

I. Introduction

MULTIPLE kernel learning (MKL) has been a topic of great interest in recent years [1]–[8]. There has been a significant amount of work on both the design of general algorithms and the application for specific domains, such as computer vision and remote sensing. For a more comprehensive review of MKLs please refer to [9]. Simply put, instead of requesting the user to choose a proper kernel, MKL seeks to learn a linear [1]–[3], [10]–[12] or nonlinear [13]–[20] combination of base kernels that optimize a given performance measure, such as an support vector machine (SVM)-type objective function [1]–[3], [10], [12], [21]–[23], kernel-target alignment [24], or Fisher's discriminant [23], [25], [26]. Based on these above, there is a large body of literature addressing various efficient solutions to the associated optimization problems. With the exception of a few publications considering a two-stage technique [11], [27] that learning a convex combination of M kernels and a kernel-based learning algorithm separately, the great majority of the techniques focus on learning a natural one-stage method, which consists of minimizing an objective function both with respect to the kernel combination parameters and the hypothesis chosen. Extensive experiments conducted in [9] suggest that while some MKL variants may be preferred for specific application, linear combination in the SVM-based one-stage framework is more reasonable for combining general purpose Gaussian kernels. Moreover, theoretically the later often presents a convex problem with efficient solving strategy and guaranteed convergence.

Hence, here we consider more specifically the problem of learning a linear combination in the SVM-based one-stage learning framework, as in much of the previous work in this area. This framework provides a flexible framework for kernel learning, but it is also too general to have a general optimal solution [1], [28], [29]. Thus, it is responsible to regularize the combined kernel weights for guaranteed optimality. Specifically, to support the interpretability and scalability, ℓ_1-norm constraint with sparse solutions in terms of the kernel weights was first introduced. Considerable progress has been made in the design of efficient algorithms [1]–[3], [21], and in the derivation of favorable theoretical guarantees [30]–[32]. Nevertheless, in practical application, ℓ_1-MKL is rarely observed to outperform even a trivial uniform combination of base kernels [13]–[20], [28]. Thus Kloft [33] founded a general MKL framework for arbitrary ℓ_p-norms ($p \geq 1$) and theoretically proved the superiority of ℓ_p-norm MKL to ℓ_1-norm MKL in a nonsparse scenario using the Rademacher complexity bound.

Despite tunable variable p ($p \geq 1$) to support controlled intrinsic sparsity, the ℓ_p-norms are "1-D" vector regularizer with just first-order linear constraints imposed on the kernel weights. Further reveal of the dependences and interactions among kernels is considered to be crucial for classification [29], [34]. In fact, for linearly combined kernels, the regularizer [35], [36] on the kernel weights can significantly impact the learning characteristics [29], [33]. In [37], they work at the level of group membership, but without

Manuscript received March 19, 2017; revised October 7, 2017; accepted December 15, 2017. Date of publication January 15, 2018; date of current version September 17, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671388, in part by the National Key Research and Development Program of China under Grant 2016YFC1400200, and in part by Fundamental Research Funds for Central Universities. *(Corresponding author: Yixin Yang.)*

Y. Han, Y. Yang, and Y. Ma are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yina.han@nwpu.edu.cn; yxyang@nwpu.edu.cn; ylma@nwpu.edu.cn).

X. Li is with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

Q. Liu is with the Institute of Navy Research of China, Beijing 100073, China (e-mail: qyliu@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2785329

quantitatively acknowledging these interactions. In [38], they impose a mixed $\ell_{1,2}$ norm regularizer, which automatically enforce the sparsity at the group feature level and learn a compact feature representation simultaneously. Recently, a growing body of work is demonstrating how the relatedness among variables naturally corresponds to learning under structural matrix constraints [39]. Various types of learning problems have involved the matrix regularization, including multiclass categorization [40], multitask and multiview learning [41], [42], online PCA [43], and latent subspace identification [44]. In [45] and [46], a matrix norm was also studied within the framework of group Lasso analysis. In our previous work [34], a matrix regularized MKL with data-dependent nonlinear combination of base kernels, referred as localized MKL, was developed. Based on the stability analysis theory, [47] further analyzed how this regularization affected the performance of the learning algorithm. Nevertheless, localized MKL generally forms a difficult nonconvex problem, and localization [48], [49] requires preclustering of the input space, which thus involves *a priori* determination of the specific clustering algorithm and the number of clusters. As mentioned above, its linear MKL counterpart is a more mainstream direction with widespread application. In [29], kernel-pair relationship in MKL is explicitly denoted by matrix $\mathbf{Q} \in \mathbb{R}^{M\times M}$, and the regularization is imposed on $\mathbf{Q}$ directly. This $\mathbf{Q}$-norm MKL has been successfully applied to neuroimaging, but an elaborate design of domain specific $\mathbf{Q}$ is required in advance.

Inspired by these works, we propose a new MKL framework with more general matrix regularization mechanism. Specifically, by organizing the kernel weights into a matrix, we present a matrix (r, p)-norm MKL $(1 \le r, p < \infty)$ with efficient optimization strategy and guaranteed convergence to encode and reveal the degree of the dependences among kernels by parameter-pair (r, p). A theoretical analysis of such a regularizer is performed using a Rademacher complexity bound, and we also prove that the generalization bounds for (r, p)-norm MKL is strictly better than that of ℓ_p-norm MKL. Finally, we report the results on seven UCI data sets to demonstrate the effectiveness of our technique.

II. From ℓ_p-MKL to (r,p)-MKL

A. ℓ_p-MKL Model

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N}$ denote a set of N labeled training samples, where $\mathbf{x}_i, i = 1,\ldots,N$, lies in some input space $\mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1,+1\}$, for binary classification. Given M different mappings $\phi_m : \mathcal{X} \to \mathcal{H}_m, m = 1,\ldots,M$, each $\mathcal{H}_m$ represents a reproducing kernel Hilbert space (RKHS) and is endowed with an associated kernel K_m. MKL approaches consider to learn a linear combination of these kernels $K = \sum \beta_m K_m$. This corresponds to taking a direct sum of the RKHS, and scaling the axes of $\mathcal{H}_m$ by $\sqrt{\beta_m}$ with $m = 1,\ldots,M$, respectively [29]. Hence, when casting into the SVM framework, the margin regularizer $(1/2)\|\mathbf{w}\|^2$ becomes a weighted sum $(1/2)\sum_{m=1}^{M}(\|\mathbf{w}_m\|^2_{\mathcal{H}_m}/\beta_m)$, and the ℓ_p-MKL$(p \ge 1)$ primal problem is given as [33]

$$\min_{\substack{\mathbf{w},b,\boldsymbol{\beta},\\ \boldsymbol{\epsilon}:\boldsymbol{\beta}\ge 0,\boldsymbol{\epsilon}\ge 0}} \frac{1}{2}\sum_{m=1}^{M}\frac{\|\mathbf{w}_m\|^2_{\mathcal{H}_m}}{\beta_m} + C\sum_{i=1}^{n}\epsilon_i + \|\boldsymbol{\beta}\|_p^2$$
$$\text{s.t. } \forall i:\ y_i\left(\sum_{m=1}^{M}\langle \mathbf{w}_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b\right) \ge 1 - \epsilon_i. \quad (1)$$

Applying Lagranges theorem, we turn problem (1) to the dual space

$$\max_{\boldsymbol{\alpha}}\ \mathbf{1}^T\boldsymbol{\alpha} - \|\mathbb{G}\|_q$$
$$\text{s.t. } \mathbf{y}^T\boldsymbol{\alpha} = 0 \text{ and } \mathbf{0} \le \boldsymbol{\alpha} \le C\mathbf{1} \quad (2)$$

where

$$\mathbb{G} = \left(\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^T\beta_m\mathbf{K}_m(\boldsymbol{\alpha}\circ\mathbf{y})\right)_{m=1}^{M} \in \mathbb{R}^M \quad (3)$$

$\circ$ denotes elementwise Hadamard product, and $\|\cdot\|_q$ represents the dual norm of the primal norm penalty $\|\boldsymbol{\beta}\|_p^2$ but imposed on the vector $\mathbb{G}$, where p and q following the identity $(1/p)+(1/q) = 1$. Note that at optimality, $\mathbf{w}_m = \beta_m(\boldsymbol{\alpha}\circ\mathbf{y})^T\phi_m(X)$, the term $\mathbb{G}_m = (\boldsymbol{\alpha}\circ\mathbf{y})^T\beta_m K_m(\boldsymbol{\alpha}\circ\mathbf{y}) = (\|\mathbf{w}_m\|^2_{\mathcal{H}_m}/\beta_m)$ becomes the vector of scaled classifier margins. Hence, the dual norm measures the margin of MKL in each RKHS [29].

B. (r,p)-MKL Model

Compared with the standard vector ℓ_p-MKL, (r, p)-MKL is featured by that the standard "1-D" vector penalty on $\boldsymbol{\beta}$, as well as the corresponding vector dual-norm penalty on $\mathbb{G}$, is substituted with a class of "2-D" matrix penalty functions, expressed as $\|\mathcal{B}\|_{r,p}$, where matrix $\mathcal{B}$ is the tensor product of $\boldsymbol{\beta}$, namely, $\mathcal{B} = \boldsymbol{\beta}\otimes\boldsymbol{\beta} = \boldsymbol{\beta}\boldsymbol{\beta}^T = (\beta_i\beta_j)_{M\times M}$. The (r, p) norms $(1 \le r, p < \infty)$ naturally generalize the ℓ_p norms to matrices by first conducting a r-norm on the columns and then a p-norm on these values

$$\|\mathcal{B}\|_{r,p} = \left(\sum_{i=1}^{N}\|\mathcal{B}^i\|_r^p\right)^{\frac{1}{p}} \quad (4)$$

where $\mathcal{B}^i$ and $\mathcal{B}_j$ denote the ith column and the jth row, respectively, of matrix $\mathcal{B}$. Reference [42, Lemma 1] has shown that the definition of (r, p) norm in (4) is indeed a norm on the space of matrices. In this framework, the burden of choosing the regularization parameter p is deferred to a choice of parameters r and p. This regularizer gives the algorithm greater flexibility while with controlled model complexity.

Considering the sequential second-order Taylor expansion, ℓ_p-norm and (r, p)-norm can be approximated by the form of

$$\|\boldsymbol{\beta}\|_p^p \approx \frac{p(p-1)}{2}\sum_{m=1}^{M}(\tilde{\beta}_m)^{p-2}\beta_m^2$$
$$- p(p-2)\sum_{m=1}^{M}(\tilde{\beta}_m)^{p-1}\beta_m + \frac{p(p-3)}{2} + 1 \quad (5)$$

and

$$\begin{aligned}\|\mathcal{B}\|_{r,p}^{rp} \approx \sum_{m=1}^{M}\sum_{m'=1}^{M}\Bigg[&\frac{p(p-1)}{2}(\tilde{\beta}_m)^{p-2}(\tilde{\beta}_{m'})^{r}\\ &+\frac{r(r-1)}{2}(\tilde{\beta}_m)^{r-2}(\tilde{\beta}_{m'})^{p}\Bigg]\beta_m^2\\ &+rp\sum_{m=1}^{M}\sum_{m'=1}^{M}(\tilde{\beta}_m)^{p-1}(\tilde{\beta}_{m'})^{r-1}\beta_m\beta_{m'}\\ &-(p+r-2)\sum_{m=1}^{M}\sum_{m'=1}^{M}\big[p(\tilde{\beta}_m)^{p-1}(\tilde{\beta}_{m'})^{r}\\ &+r(\tilde{\beta}_m)^{r-1}(\tilde{\beta}_{m'})^{p}\big]\beta_m\\ &+\frac{(p+r)(p+r-3)}{2}+1. \qquad (6)\end{aligned}$$

Equations (5) and (6) offer an intuitive view of what ℓ_p-MKL and (*r*, *p*)-MKL are actually optimizing. Compared with 1-D ℓ_p-norm constraint over the kernel weights, the proposed 2-D (*r*, *p*)-norm constraint explores the interactions of kernels by adding cross terms between different kernels in the regularizer.

The model we optimize is

$$\begin{aligned}\min_{\substack{\mathbf{w},b,\boldsymbol{\beta},\\ \boldsymbol{\epsilon}:\boldsymbol{\beta}\geq 0,\boldsymbol{\epsilon}\geq 0}} \ &\frac{1}{2}\sum_{m=1}^{M}\frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m}+C\sum_{i=1}^{N}\epsilon_i+\|\mathcal{B}\|_{r,p}\\ \text{s.t. } \forall i:\ &y_i\left(\sum_{m=1}^{M}\langle\mathbf{w}_m,\phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m}+b\right)\geq 1-\epsilon_i. \qquad (7)\end{aligned}$$

Using the fact that the dual of (*r*, *p*)-norm is the (*s*, *q*)-norm, where *s* and *q* are the exponents dual to *r* and *p*, respectively, i.e., $(1/r)+(1/s)=1$ and $(1/p)+(1/q)=1$ [50]. The dual problem becomes (see also the Appendix for mathematical details about the derivation)

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\ \mathbf{1}^T\boldsymbol{\alpha}-\|\mathbb{G}\mathbb{G}^T\|_{s,q}^{\frac{1}{2}}\\ &\text{s.t. } \mathbf{y}^T\boldsymbol{\alpha}=0 \text{ and } \mathbf{0}\leq\boldsymbol{\alpha}\leq C\mathbf{1}. \qquad (8)\end{aligned}$$

It is easy to see that $\|\mathbb{G}\mathbb{G}^T\|_{s,q}=\|\mathbb{G}\|_s\|\mathbb{G}\|_q$, and if $r=p$, the vector form of (2), i.e., ℓ_p-MKL, becomes a special case of (*r*, *p*)-MKL, because $\|\mathcal{B}\|_{p,p}=\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{p,p}=\|\boldsymbol{\beta}\|_p^2$.

III. Optimization Strategies

As most MKL solvers [1]–[3], [10], [12], [21] do, the optimization variables of the primal problem (7) are divided into two groups, which are $(\mathbf{w}, b)$ and $\boldsymbol{\beta}$, respectively. Then, a two-layer alternating optimization over each group of variables while holding the other group at their most recent values is performed.

A. Fixing $\boldsymbol{\beta}$, Optimizing (w, b)

At optimality, we have

$$\mathbf{w}_m=\beta_m(\boldsymbol{\alpha}\circ\mathbf{y})^T\phi_m(X) \qquad (9)$$

and hence

$$\|\mathbf{w}_m\|_{\mathcal{H}_m}^2=\beta_m^2(\boldsymbol{\alpha}\circ\mathbf{y})^T K_m(\boldsymbol{\alpha}\circ\mathbf{y}) \qquad (10)$$

where $\boldsymbol{\alpha}$ can be conveniently obtained using any off-the-shelf SVM solver with the combined kernel $\sum_{m=1}^{M}\beta_m\mathbf{K}_m$.

B. Fixing (w, b), Optimizing $\boldsymbol{\beta}$

Proposition 1: Let $r, p > 1$, with the optimal $\mathbf{w}\neq 0$ and b from the above procedure, $\forall m=1,\ldots,M$ the minimal β_m of problem (7) can be expressed as

$$\beta_m=\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^2}{\left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_i}\|_{\mathcal{H}_i}^{2p}\right)^{\frac{1}{2p}}\left(\sum_{j=1}^{M}\|\mathbf{w}_{\beta_j}\|_{\mathcal{H}_j}^{2r}\right)^{\frac{1}{2r}}} \qquad (11)$$

or

$$\beta_m=\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^2}{\left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_i}\|_{\mathcal{H}_i}^{2p}\right)^{\frac{1}{p+r}}\left(\sum_{j=1}^{M}\|\mathbf{w}_{\beta_j}\|_{\mathcal{H}_j}^{2r}\right)^{\frac{1}{p+r}}} \qquad (12)$$

where

$$\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^2=\frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\left(\frac{\beta_m}{\|\boldsymbol{\beta}\|_p}\right)^p+\left(\frac{\beta_m}{\|\boldsymbol{\beta}\|_r}\right)^r}. \qquad (13)$$

Proof: Optimization Problem (7) can be equivalently translated into [33]

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\beta}:\boldsymbol{\beta}\geq 0}\ &\tilde{C}\sum_{i=1}^{n}\ell\left(\sum_{m=1}^{M}\langle\mathbf{w}_m,\phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m}+b, y_i\right)\\ &+\frac{1}{2}\sum_{m=1}^{M}\frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m}+\frac{\mu}{2}\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} \qquad (14)\end{aligned}$$

where $\mu > 0$, and ℓ is a convex loss function. With the optimal $(\mathbf{w}, b)$ from the above procedure, we set the gradient of the above objective with respect to $\boldsymbol{\beta}$ to zero, yielding the following optimal condition of $\boldsymbol{\beta}$:

$$-\frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{2\beta_m^2}+\frac{\mu}{2}\cdot\frac{\partial(\|\mathcal{B}\|_{r,p})}{\partial\beta_m}=0,\quad \forall m=1,\ldots,M. \qquad (15)$$

The first derivative of $\|\mathcal{B}\|_{r,p}$ with respect to β_m, $m=1,\ldots,M$ can be expressed as

$$\begin{aligned}\frac{\partial(\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p})}{\partial\beta_m}&=\frac{\partial(\|\boldsymbol{\beta}\|_r\|\boldsymbol{\beta}\|_p)}{\partial\beta_m}\\ &=\beta_m^{p-1}\|\boldsymbol{\beta}\|_p^{1-p}\|\boldsymbol{\beta}\|_r+\beta_m^{r-1}\|\boldsymbol{\beta}\|_r^{1-r}\|\boldsymbol{\beta}\|_p\\ &=\frac{\beta_m^{p-1}}{\|\boldsymbol{\beta}\|_p^p}\|\boldsymbol{\beta}\|_p\|\boldsymbol{\beta}\|_r+\frac{\beta_m^{r-1}}{\|\boldsymbol{\beta}\|_r^r}\|\boldsymbol{\beta}\|_r\|\boldsymbol{\beta}\|_p. \qquad (16)\end{aligned}$$

According to [33, Th. 1], the regularization in (14) known as Tikhonov regularization is equivalent to the so-called Ivanov regularization in the following form:

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\beta}:\boldsymbol{\beta}\geq 0}\ &\tilde{C}\sum_{i=1}^{n}\ell\left(\sum_{m=1}^{M}\langle\mathbf{w}_m,\phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m}+b, y_i\right)\\ &+\frac{1}{2}\sum_{m=1}^{M}\frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m}\\ \text{s.t. } &\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p}\leq 1. \qquad (17)\end{aligned}$$

Hence for an optimal $\boldsymbol{\beta}$, the constraint $\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p}\leq 1$ in (17) is at the upper bound, that is $\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p}=1$, and equivalently

$\|\boldsymbol{\beta}\|_r\|\boldsymbol{\beta}\|_p = 1$ holds. Hence, (16) is reduced to the following form:

$$\frac{\partial(\|\mathcal{B}\|_{r,p})}{\partial\beta_m} = \frac{\beta_m^{p-1}}{\|\boldsymbol{\beta}\|_p^p} + \frac{\beta_m^{r-1}}{\|\boldsymbol{\beta}\|_r^r}. \tag{18}$$

Plugging (18) into (15), and we define a new variable $\mathbf{w}_{\beta_m}$

$$\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^2 = \frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\left(\frac{\beta_m}{\|\boldsymbol{\beta}\|_p}\right)^p + \left(\frac{\beta_m}{\|\boldsymbol{\beta}\|_r}\right)^r}. \tag{19}$$

Then (15) translates into the following optimality condition:

$$\forall m = 1, \ldots, M\colon\ \beta_m = \frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^2}{\mu}. \tag{20}$$

Again, $\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} = 1$ is equivalent to $\|\boldsymbol{\beta}\|_r\|\boldsymbol{\beta}\|_p = 1$. Substituting (20) into this upper bound, namely

$$\left(\sum_{i=1}^{M}\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2p}}{\mu^p}\right)^{\frac{1}{p}}\left(\sum_{i=1}^{M}\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2r}}{\mu^r}\right)^{\frac{1}{r}} = 1 \tag{21}$$

yields

$$\mu = \left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2p}\right)^{\frac{1}{2p}}\left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2r}\right)^{\frac{1}{2r}}. \tag{22}$$

Then substituting (22) into (20), we have the claimed formula (11).

Moreover, $\|\boldsymbol{\beta}\|_r\|\boldsymbol{\beta}\|_p = 1$ is also equivalent to $\|\boldsymbol{\beta}\|_r^r\|\boldsymbol{\beta}\|_p^p = 1$. Similarly, substituting (20) into this upper bound equation, namely

$$\left(\sum_{i=1}^{M}\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2p}}{\mu^p}\right)\left(\sum_{i=1}^{M}\frac{\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2r}}{\mu^r}\right) = 1 \tag{23}$$

yields

$$\mu = \left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2p}\right)^{\frac{1}{p+r}}\left(\sum_{i=1}^{M}\|\mathbf{w}_{\beta_m}\|_{\mathcal{H}_m}^{2r}\right)^{\frac{1}{p+r}} \tag{24}$$

and substituting (24) into (20), we have the claimed formula (12). ■

It is clear that the introduction of $\|\mathbf{w}_{\beta_m}\|^2$ in optimizing $\boldsymbol{\beta}$ introduces more interactions between the kernel weights and makes them more involved during optimization than ℓ_p-MKL does. Hence, (r, p)-MKL requires alternative solution mechanisms to update $\|\mathbf{w}_{\beta_m}\|^2$ with the current estimation of β_m according to (13).

We now summarize the alternating optimization for (r, p)-norm MKL training discussed above in Algorithm 1.

Algorithm 1 Simple (r, p)-Norm MKL Training Algorithm Using Alternating Between the Analytical Updates of $\boldsymbol{\beta}$ and the SVM Computations

```
Initialization: feasible $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
repeat
    Calculate $\boldsymbol{\alpha}$ according to Equation (8) (e.g., SVM)
    Calculate $\|\mathbf{w}_m\|^2$ for all $m = 1, \ldots, M$ according to Equation (10)
    Calculate $\|\mathbf{w}_{\beta_m}\|^2$ for all $m = 1, \ldots, M$ according to Equation (13)
    Update $\boldsymbol{\beta}$ according to Equation (11) or Equation (12),
until Convergence
```

C. Convergence

Since the regularization term in (14) can be expressed as [50]

$$\begin{aligned}\|\mathcal{B}\|_{r,p} &= \sup\{\|\mathcal{B}\mathbf{v}\|_r \big| \|\mathbf{v}\|_p = 1\}\\ &= \sup\{\mathbf{u}^T\mathcal{B}\mathbf{v} \big| \|\mathbf{u}\|_{r*}\|\mathbf{v}\|_p = 1\}\end{aligned} \tag{25}$$

where we use the fact that

$$\|\mathbf{z}\|_p = \sup\{\mathbf{u}^T\mathbf{z} \big| \|\mathbf{u}\|_{p*} = 1\}. \tag{26}$$

Since $\|\mathcal{B}\|_{r,p}$ can be expressed as a supremum of linear functions of $\mathcal{B}$, it is a convex function, and our (r, p)-MKL mode (14) is a convex optimization problem. Since we have shown that our formulation can be precisely optimized at each step, it can be solved optimally.

IV. Theoretical Analysis

This section presents a theoretical analysis of (r, p)-norm MKL based on the established theory of Rademacher complexities. Following [51], we define the following hypothesis class for $r, p \in [1, \infty)$:

$$H_M^{r,p} := \left\{h : \mathcal{X} \to \mathbb{R} \big| h(\mathbf{x}) = \sum_{m=1}^{M}\sqrt{\beta_m}\langle\mathbf{w}_m, \phi_m(\mathbf{x})\rangle_{\mathcal{H}_m},\right.$$
$$\left.\|\mathbf{w}\|_{2,r} \le D, \|\mathbf{w}\|_{2,p} \le D, \|\mathcal{B}\|_{r,p} \le 1\right\}. \tag{27}$$

Solving the primal problem of (r, p)-MKL (7) corresponds to minimizing the empirical risk in the above hypothesis class. To further investigate the bound of the generalization error of the above class with respect to an i.i.d. sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \in \mathcal{X} \times \{-1, 1\}$ from an arbitrary distribution P, we resort to the global Rademacher complexity

$$\mathcal{R}\big(H_M^{r,p}\big) := \mathbb{E}\left[\sup_{h \in H_M^{r,p}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i h(\mathbf{x}_i)\right] \tag{28}$$

where $\sigma_1, \ldots, \sigma_N$ are independent Rademacher variables, which take the value -1 or $+1$ with the same probability of 0.5, and $\mathbb{E}$ is the expectation operator on all random variables, that is, $\sigma_i, \mathbf{x}_i$, and $y_i (i = 1, \ldots, N)$. With the Rademacher complexity, plenty of results are available for bounding the generalization error [52], [53].

For ℓ_p-MKL, the Rademacher complexity has been proven in [54].

Theorem 1 (Global Rademacher Complexities of ℓ_p-MKL [54]): Assume the kernels are uniformly bounded, that is, $\|k\|_\infty \le B \le \infty$, almost surely. Then for any sample of size N, any $M > 1$ and $p \geq 1$, the global Rademacher complexity of the multikernel class H_M^p can be bounded as

$$\forall t \geq p:\ \mathcal{R}\big(H_M^p\big) \leq Dt^* \sqrt{\frac{e}{N} \big\| (\mathrm{tr}(\mathbf{J}_m))_{m=1}^M \big\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be} D M^{\frac{1}{t^*}} t^*}{N} \tag{29}$$

where $\mathrm{tr}(\mathbf{J}_m) = \mathbb{E}((1/N)\sum_{i=1}^N \|\phi_m(\mathbf{x}_i)\|_2^2)$.

We then show a simple ℓ_p-norm to (r, p)-norm conversion technique for the Rademacher complexity of (r, p)-MKL.

Theorem 2 (ℓ_p-to-(r,p) Conversion): For any sample of size N, any $M > 1$ and $1 \leq r, p \leq \infty$, the Rademacher complexity of the hypothesis set $H_M^{r,p}$ can be bounded in terms of H_M^p and H_M^r

$$\mathcal{R}\big(H_M^{r,p}\big) \leq \sqrt{M^{\frac{1}{2r} - \frac{1}{2p}}} \mathcal{R}\big(H_M^p\big) \tag{30}$$

and

$$\mathcal{R}\big(H_M^{r,p}\big) \leq \sqrt{M^{\frac{1}{2p} - \frac{1}{2r}}} \mathcal{R}\big(H_M^r\big). \tag{31}$$

Proof: By Hölder's inequality [55], denoting $\boldsymbol{\beta}^p := (\beta_1^p, \ldots, \beta_M^p)^T$, we have for all nonnegative $\boldsymbol{\beta} \in \mathbb{R}^M$

$$\begin{aligned} \|\boldsymbol{\beta}\|_p &= (\mathbf{1}^T \boldsymbol{\beta}^p)^{\frac{1}{p}} \leq \big(\|\mathbf{1}\|_{(\frac{r}{p})^*} \|\boldsymbol{\beta}^p\|_{\frac{r}{p}}\big)^{\frac{1}{p}} \\ &= M^{\frac{1}{p(r/p)^*}} \|\boldsymbol{\beta}\|_r = M^{\frac{1}{p} - \frac{1}{r}} \|\boldsymbol{\beta}\|_r \end{aligned} \tag{32}$$

and with

$$\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} = \|\boldsymbol{\beta}\|_r \|\boldsymbol{\beta}\|_p \tag{33}$$

we have

$$\|\boldsymbol{\beta}\|_p \leq \big(M^{\frac{1}{p} - \frac{1}{r}} \|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p}\big)^{\frac{1}{2}}. \tag{34}$$

Hence

$$\begin{aligned} &\mathcal{R}\big(H_M^{r,p}\big) \\ &\overset{\text{Def.}}{=} \mathbb{E}\left[\sup_{\mathbf{w},\boldsymbol{\beta}:\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} \leq 1} \frac{1}{N} \sum_{i=1}^N \sigma_i \sum_{m=1}^M \sqrt{\beta_m} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m}\right] \\ &\leq \mathbb{E}\left[\sup_{\mathbf{w},\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_p \leq M^{\frac{1}{2p} - \frac{1}{2r}}} \frac{1}{N} \sum_{i=1}^N \sigma_i \sum_{m=1}^M \sqrt{\beta_m} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m}\right] \\ &= \mathbb{E}\left[\sup_{\mathbf{w},\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_p \leq 1} \frac{1}{N} \sum_{i=1}^N \sigma_i \sum_{m=1}^M \sqrt{\beta_m M^{\frac{1}{2r} - \frac{1}{2p}}} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m}\right] \\ &\overset{\text{Def.}}{=} \sqrt{M^{\frac{1}{2r} - \frac{1}{2p}}} \mathcal{R}\big(H_M^p\big). \end{aligned} \tag{35}$$

Similarly, we also have

$$\mathcal{R}\big(H_M^{r,p}\big) \leq \sqrt{M^{\frac{1}{2p} - \frac{1}{2r}}} \mathcal{R}\big(H_M^r\big). \tag{36}$$

■

With the above result, we can make use of any existing bound on the Rademacher complexity of H_M^p (or H_M^r) in order to obtain a generalization bound for $H_M^{r,p}$. Specifically, based on theory IV, the Rademacher complexities of the hypothesis classes $H_M^{r,p}$ can be bounded as follows.

TABLE I

NUMBER OF MULTIPLE KERNELS M USED IN EACH OF THE SEVEN UCI DATA SETS

dataset	M
ionosphere	33
heart	13
sonar	60
pima	8
german	24
liverdisorder	6
wdbc	30

Corollary 1 (Global Rademacher Complexities of (r,p)-MKL): Assume the kernels are uniformly bounded, that is, $\|k\|_\infty \leq B \leq \infty$, almost surely. Then for any sample of size N, any $M > 1$ and $1 \leq r, p \leq \infty$, the global Rademacher complexity of the multikernel class $H_M^{r,p}$ can be bounded as

$$\mathcal{R}(H_M^{r,p}) \leq Dt_1^* \sqrt{\frac{eM^{\frac{1}{2t_1^*} - \frac{1}{2t_2^*}}}{N} \big\| (\mathrm{tr}(\mathbf{J}_m))_{m=1}^M \big\|_{\frac{t_1^*}{2}}} + \frac{\sqrt{Be} D M^{\frac{5}{4t_1^*} - \frac{1}{4t_2^*}} t_1^*}{N} \tag{37}$$

where $t_1 = \min(r, p)$, $t_2 = \max(r, p)$, and $\mathrm{tr}(\mathbf{J}_m) = \mathbb{E}((1/N)\sum_{i=1}^N \|\phi_m(\mathbf{x}_i)\|_2^2)$.

Proof: Since for all $t \geq p$, $\mathcal{R}(H_M^p) \leq \mathcal{R}(H_M^t)$ holds [54]. $\mathcal{R}(H_M^r)$ and $\mathcal{R}(H_M^p)$ can be compactly bounded by substituting the variable t in (29) with r and p, respectively

$$\begin{aligned} \mathcal{R}\big(H_M^r\big) &\leq Dr^* \sqrt{\frac{e}{N} \big\| (\mathrm{tr}(\mathbf{J}_m))_{m=1}^M \big\|_{\frac{r^*}{2}}} + \frac{\sqrt{Be} D M^{\frac{1}{r^*}} r^*}{N} \\ \mathcal{R}\big(H_M^p\big) &\leq Dp^* \sqrt{\frac{e}{N} \big\| (\mathrm{tr}(\mathbf{J}_m))_{m=1}^M \big\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be} D M^{\frac{1}{p^*}} p^*}{N}. \end{aligned} \tag{38}$$

Similarly, if $r < p$, we have $\mathcal{R}(H_M^r) \leq \mathcal{R}(H_M^p)$, and vice versa. Then, according to (35) and (36), $\mathcal{R}(H_M^{r,p})$ can be compactly bounded as

$$\mathcal{R}\big(H_M^{r,p}\big) \leq \sqrt{M^{-|\frac{1}{2r} - \frac{1}{2p}|}} \mathcal{R}\big(H_M^{\min(r,p)}\big). \tag{39}$$

Plugging (38) into (39), and let $t_1 = \min(r, p)$, $t_2 = \max(r, p)$ leads exactly the statement of the corollary. ■

Corollary 2: Assume the kernels are uniformly bounded, that is, $\|k\|_\infty \leq B \leq \infty$, almost surely. Then for any sample of size N, any $M > 1$ and $1 \leq r, p \leq \infty$, the Rademacher complexity bound for the matrix (r, p)-norm hypothesis class $H_M^{r,p}$ is strictly lower than that for the vector ℓ_p-norm hypothesis class H_M^p and for the ℓ_r-norm hypothesis class H_M^r.

Proof: Again since for all $r \leq p$, $\mathcal{R}(H_M^r) \leq \mathcal{R}(H_M^p)$ holds. We have

$$\mathcal{R}\big(H_M^{r,p}\big) \leq \sqrt{M^{\frac{1}{2p} - \frac{1}{2r}}} \mathcal{R}(H_M^r) \leq \mathcal{R}\big(H_M^r\big) \leq \mathcal{R}\big(H_M^p\big). \tag{40}$$

The reverse is also true. This just concludes the proof of this corollary. ■

TABLE II

Comparison of Three Implementations of (r, p)-MKL With Their Optimal C, r, and p Values Over Each of the Seven UCI Data Sets. (a) Test Accuracies (%). (b) Training Time (Seconds). (c) Support Vector Percentages (%)

(a) Test accuracies (%)

dataset	(r,p)-MKL0	(r,p)-MKL1	(r,p)-MKL2	(r,p)-MKL3
ionosphere	93.31 ± 2.19	92.39 ± 2.52	93.94 ± 2.15	94.08 ± 2.02
heart	85.65 ± 4.76	86.76 ± 4.00	84.07 ± 2.51	84.35 ± 2.44
sonar	84.52 ± 4.34	84.64 ± 3.58	84.64 ± 4.05	84.64 ± 4.05
pima	76.75 ± 2.59	76.49 ± 3.00	76.88 ± 3.75	76.88 ± 3.75
german	75.98 ± 2.61	75.63 ± 1.80	76.10 ± 2.19	76.10 ± 2.19
liverdisorder	72.32 ± 4.84	73.41 ± 4.87	74.64 ± 5.35	74.28 ± 5.23
wdbc	96.80 ± 1.22	97.32 ± 1.38	97.37 ± 1.42	97.59 ± 1.24
Direct Comparison	6-0-1	5-1-1	3-3-1	–
t-Test	0-6-1	1-5-1	0-7-0	–
Signed Rank	Tie	Tie	Tie	–

(b) Training time (sec.)

dataset	(r,p)-MKL0	(r,p)-MKL1	(r,p)-MKL2	(r,p)-MKL3
ionosphere	0.14 ± 0.00	0.16 ± 0.01	0.15 ± 0.00	0.14 ± 0.01
heart	0.11 ± 0.00	0.12 ± 0.00	0.09 ± 0.00	0.12 ± 0.00
sonar	0.14 ± 0.00	0.13 ± 0.01	0.10 ± 0.00	0.10 ± 0.00
pima	0.34 ± 0.03	0.53 ± 0.01	0.33 ± 0.03	0.33 ± 0.03
german	0.67 ± 0.02	1.08 ± 0.04	1.01 ± 0.07	0.57 ± 0.01
liverdisorder	0.15 ± 0.01	0.13 ± 0.01	0.13 ± 0.01	0.13 ± 0.01
wdbc	0.32 ± 0.00	0.41 ± 0.03	0.28 ± 0.02	0.31 ± 0.00
Direct Comparison	5-1-1	5-2-0	2-3-2	–
t-Test	4-1-2	4-1-2	2-3-2	–
Signed Rank	Tie	Tie	Tie	–

(c) Support vector percentages (%)

dataset	(r,p)-MKL0	(r,p)-MKL1	(r,p)-MKL2	(r,p)-MKL3
ionosphere	45.18 ± 1.25	36.25 ± 0.81	34.16 ± 0.75	38.25 ± 0.79
heart	46.90 ± 1.77	43.84 ± 1.22	42.94 ± 0.99	44.17 ± 1.21
sonar	43.40 ± 2.41	51.08 ± 1.47	49.70 ± 1.71	49.70 ± 1.71
pima	51.26 ± 1.86	49.40 ± 1.11	55.35 ± 1.19	55.35 ± 1.19
german	53.02 ± 0.94	56.05 ± 0.35	57.14 ± 0.36	57.14 ± 0.36
liverdisorder	64.57 ± 2.00	70.56 ± 1.06	70.20 ± 0.77	70.89 ± 1.05
wdbc	10.23 ± 0.92	9.67 ± 0.68	14.67 ± 0.79	12.57 ± 0.66
Direct Comparison	2-0-5	1-0-6	1-3-3	–
t-Test	2-2-3	2-1-4	1-3-3	–
Signed Rank	Tie	Tie	Tie	–

V. Experiments

This section evaluates the proposed (r, p)-MKL on seven UCI data sets, which include a variety of classification tasks and are frequently used in MKL experiments [3], [12], [56]. For UCI data sets, each dimension records a specific attribute. Specifically, in the data set of "sonar," it is a specific angle at which sonar signals bouncing off an object. Hence, we just construct a certain kernel for each dimension, and the behavior of regularization strategy imposed on the kernel weights can help to reveal the contribution or the interactional relationship of associated attributes in the process of classification. The number of multiple kernels used in each data set is listed in Table I, which is equal to the number of data dimensions. We adopt three commonly used kernels: linear kernel (K_L), polynomial kernel (K_P), and Gaussian kernel (K_G)

$$\begin{aligned} K_L(\mathbf{x}_i, \mathbf{x}_j) &= \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ K_P(\mathbf{x}_i, \mathbf{x}_j) &= (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^q \\ K_G(\mathbf{x}_i, \mathbf{x}_j) &= \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right) \end{aligned} \tag{41}$$

to construct multiple linear kernels, multiple polynomial kernels, and multiple Gaussian kernels, respectively, for each data set. We use the third degree ($q = 3$) for polynomial kernels and set the width (σ^2) to half the mean squared distance of associated dimension for Gaussian kernels.

TABLE III

Optimized p and (r, p) Values in ℓ_p-MKL and (r, p)-MKL, Respectively, With Respect to Multiple Linear, Polynomial, and Gaussian Kernels for Each of the Seven UCI Data Sets

datasets	p			(r,p)		
	K_L	K_P	K_G	K_L	K_P	K_G
ionosphere	10	10	2	$(3,4)$	$(2,4)$	$(1,5)$
heart	3	4	2	$(2,3)$	$(2,4)$	$(3,3)$
sonar	10	10	4	$(1,2)$	$(1,10)$	$(1,4)$
pima	4	4	10	$(1,10)$	$(2,4)$	$(1,6)$
german	3	10	4	$(2,3)$	$(2,4)$	$(3,3)$
liverdisorder	2	1	4	$(2,4)$	$(3,4)$	$(2,5)$
wdbc	4	10	4	$(1,4)$	$(3,4)$	$(4,5)$

Given a data set, we randomly split it into 80% for training and 20% for test. The training data are normalized to have zero mean and unit variance, and the test data are then normalized using the mean and variance of the training data. The involved parameters are the regularization parameters C and (r, p). Then for each split, we test (r, p)-MKL over a grid of values:

TABLE IV

COMPARISON OF ℓ_p-MKL AND (r, p)-MKL WITH THEIR OPTIMAL SETTINGS OVER EACH OF THE SEVEN UCI DATA SETS. (a) MKL WITH MULTIPLE K_L VALUES. (b) MKL WITH MULTIPLE K_P VALUES. (c) MKL WITH MULTIPLE K_G VALUES

(a) MKL with multiple K_L

datasets	ℓ_p-MKL				(r,p)-MKL			
	ACC (%)	Time (sec.)	SV (%)	# of Iter	ACC (%)	Time (sec.)	SV (%)	# of Iter
ionosphere	87.67 ± 2.76	$\mathbf{0.14 \pm 0.00}$	40.09 ± 1.71	2	$\mathbf{89.58 \pm 3.18}$	0.17 ± 0.00	$\mathbf{24.02 \pm 1.84}$	2
heart	85.56 ± 4.64	$\mathbf{0.09 \pm 0.00}$	50.83 ± 1.81	2	$\mathbf{86.39 \pm 4.13}$	0.11 ± 0.00	$\mathbf{46.97 \pm 1.20}$	2
sonar	$\mathbf{76.90 \pm 3.95}$	0.11 ± 0.00	$\mathbf{56.14 \pm 2.57}$	2	76.67 ± 5.04	$\mathbf{0.08 \pm 0.00}$	68.55 ± 2.09	**1**
pima	77.24 ± 3.07	$\mathbf{0.28 \pm 0.01}$	61.93 ± 0.87	2	$\mathbf{77.69 \pm 3.97}$	0.29 ± 0.01	$\mathbf{60.75 \pm 1.01}$	**1**
german	72.83 ± 17.23	$\mathbf{0.60 \pm 0.05}$	54.89 ± 1.49	2	$\mathbf{76.68 \pm 1.87}$	0.66 ± 0.01	$\mathbf{52.86 \pm 0.97}$	2
liverdisorder	69.13 ± 5.59	$\mathbf{0.11 \pm 0.00}$	80.09 ± 1.35	2	$\mathbf{69.42 \pm 5.34}$	0.12 ± 0.00	$\mathbf{79.49 \pm 0.59}$	2
wdbc	97.28 ± 0.98	$\mathbf{0.13 \pm 0.00}$	$\mathbf{11.27 \pm 0.61}$	2	$\mathbf{97.89 \pm 1.43}$	0.21 ± 0.00	11.63 ± 0.66	**1**
Direct Comparison	6-0-1	1-0-6	5-0-2	3-4-0	–	–	–	–
t-Test	3-4-0	1-3-3	3-2-2	3-4-0	–	–	–	–
Signed Rank	Win	Tie	Tie	Tie	–	–	–	–

(b) MKL with multiple K_P

datasets	ℓ_p-MKL				(r,p)-MKL			
	ACC (%)	Time (sec.)	SV (%)	# of Iter	ACC (%)	Time (sec.)	SV (%)	# of Iter
ionosphere	91.48 ± 2.68	$\mathbf{0.12 \pm 0.00}$	$\mathbf{26.75 \pm 1.39}$	2	$\mathbf{93.59 \pm 2.17}$	0.16 ± 0.00	28.21 ± 1.10	2
heart	$\mathbf{85.28 \pm 4.55}$	$\mathbf{0.08 \pm 0.00}$	43.06 ± 2.10	2	84.44 ± 5.01	0.13 ± 0.00	$\mathbf{42.70 \pm 1.81}$	2
sonar	83.10 ± 4.15	0.12 ± 0.00	$\mathbf{48.37 \pm 2.19}$	2	$\mathbf{83.33 \pm 5.93}$	0.12 ± 0.00	48.40 ± 2.59	**1**
pima	75.84 ± 2.34	$\mathbf{0.29 \pm 0.01}$	58.41 ± 1.34	2	$\mathbf{76.82 \pm 3.95}$	0.36 ± 0.01	$\mathbf{50.63 \pm 1.52}$	2
german	75.75 ± 2.06	$\mathbf{0.60 \pm 0.12}$	55.89 ± 0.77	2	$\mathbf{76.65 \pm 2.51}$	0.73 ± 0.03	$\mathbf{51.67 \pm 0.94}$	2
liverdisorder	58.91 ± 1.35	$\mathbf{0.12 \pm 0.01}$	84.33 ± 0.68	2	$\mathbf{69.28 \pm 5.96}$	0.17 ± 0.02	$\mathbf{73.15 \pm 6.77}$	2
wdbc	95.48 ± 1.15	$\mathbf{0.22 \pm 0.02}$	21.11 ± 0.86	2	$\mathbf{96.49 \pm 1.82}$	0.31 ± 0.00	$\mathbf{16.94 \pm 0.79}$	2
Direct Comparison	6-0-1	0-1-6	5-0-2	1-6-0	–	–	–	–
t-Test	3-4-0	2-2-3	3-2-2	1-6-0	–	–	–	–
Signed Rank	Win	Tie	Tie	Tie	–	–	–	–

(c) MKL with multiple K_G

datasets	ℓ_p-MKL				(r,p)-MKL			
	ACC (%)	Time (sec.)	SV (%)	# of Iter	ACC (%)	Time (sec.)	SV (%)	# of Iter
ionosphere	92.11 ± 2.52	$\mathbf{0.13 \pm 0.00}$	$\mathbf{36.46 \pm 1.43}$	3	$\mathbf{94.08 \pm 2.02}$	0.14 ± 0.01	38.25 ± 0.79	**2**
heart	86.02 ± 4.05	$\mathbf{0.07 \pm 0.00}$	45.07 ± 1.75	2	$\mathbf{86.76 \pm 4.00}$	0.12 ± 0.00	$\mathbf{43.84 \pm 1.22}$	2
sonar	$\mathbf{85.24 \pm 4.41}$	0.12 ± 0.01	50.48 ± 1.92	2	84.64 ± 4.05	$\mathbf{0.10 \pm 0.00}$	$\mathbf{49.70 \pm 1.71}$	**1**
pima	76.79 ± 2.50	0.36 ± 0.02	55.83 ± 1.12	2	$\mathbf{76.88 \pm 3.75}$	$\mathbf{0.33 \pm 0.03}$	$\mathbf{55.35 \pm 1.19}$	**1**
german	71.93 ± 17.07	0.66 ± 0.15	58.62 ± 0.92	2	$\mathbf{76.10 \pm 2.19}$	$\mathbf{0.57 \pm 0.01}$	$\mathbf{57.14 \pm 0.36}$	2
liverdisorder	71.88 ± 3.62	0.23 ± 0.27	80.83 ± 1.00	2	$\mathbf{74.64 \pm 5.35}$	$\mathbf{0.13 \pm 0.01}$	$\mathbf{70.20 \pm 0.77}$	2
wdbc	96.80 ± 0.96	$\mathbf{0.16 \pm 0.00}$	16.23 ± 0.84	2	$\mathbf{97.59 \pm 1.24}$	0.31 ± 0.00	$\mathbf{12.57 \pm 0.66}$	2
Direct Comparison	6-0-1	4-0-3	6-0-1	3-4-0	–	–	–	–
t-Test	3-4-0	3-1-3	3-3-1	3-4-0	–	–	–	–
Signed Rank	Win	Tie	Tie	Tie	–	–	–	–

$C \in \{0.01, 0.1, 1, 10, 100\}$ and $r, p \in \{1, 2, 4, 10\}$ jointly. This process is repeated 20 times, and the results are recorded in the form of the mean and standard deviation of each individual one. The finally reported results are those having the optimal average test accuracies. The experiments are implemented in MATLAB, and run on a 3.30-GHz Intel (R) Core (TM) i5-2500K CPU with 8 GB of RAM PC.

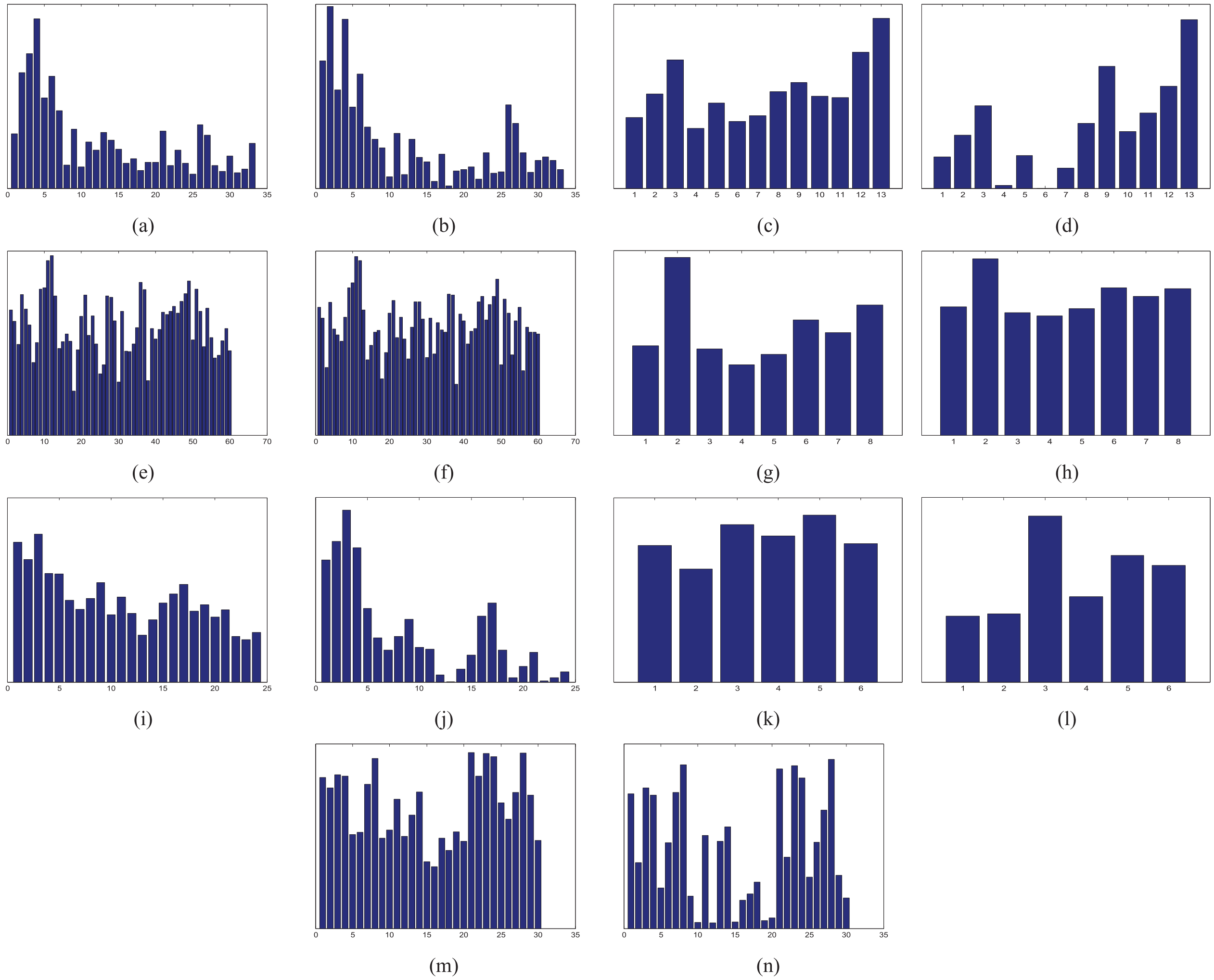
For statistically reliable results rather than would be expected by chance, we use three kinds of statistical tests, namely, direct comparison, t-test, and Wilcoxons signed rank test (at the significance level of 0.05). In the direct comparison and t-test, we use W-T-L to record the counts of Win–Tie–Loss on the seven benchmark machine learning data sets. In Wilcoxons signed rank test, Win means statistically significant superiority, Tie means no statistically significant difference between the two compared methods, and Loss means opposite results to Win.

A. Initialization of (r,p)*-MKL*

For the proposed (r, p)-MKL, we first try out four initialization strategies with multiple Gaussian kernels.

1) *(r, p)-MKL0:* Initializing $\boldsymbol{\beta}$ randomly and updating $\boldsymbol{\beta}$ with (12).
2) *(r, p)-MKL1:* Initializing $\boldsymbol{\beta}$ with $1/M$ and updating $\boldsymbol{\beta}$ with (11).
3) *(r, p)-MKL2:* Initializing $\boldsymbol{\beta}$ with ℓ_p-MKL and updating $\boldsymbol{\beta}$ with (11).
4) *(r, p)-MKL3:* Initializing $\boldsymbol{\beta}$ with ℓ_p-MKL and updating $\boldsymbol{\beta}$ with (12).

The p value of ℓ_p-MKL used for initializing (r, p)-MKL2 and (r, p)-MKL3 is just the same p value as in associated (r, p)-MKL2 and (r, p)-MKL3. Table II summarizes the average testing accuracies, training time, and support vector percentages for the aforementioned three implementation strategies. In terms of testing accuracy, (r, p)-MKL3 obtains 6-0-1 against (r, p)-MKL0 using direct comparison, that is to say, (r, p)-MKL3 outperforms (r, p)-MKL0 on six out of seven data sets and inferior to (r, p)-MKL1 on one out of seven data sets. When compared with (r, p)-MKL1, (r, p)-MKL3 obtains 5-1-1 against (r, p)-MKL1 using direct comparison, that is to say, (r, p)-MKL3 outperforms (r, p)-MKL1 on five



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n)

Fig. 1. Examples of the kernel weights distribution selected from the seven UCI data sets for ℓ_p-MKL and (r, p)-MKL. (a) Ionosphere with ℓ_p-norm. (b) Ionosphere with (r, p)-norm. (c) Heart with ℓ_p-norm. (d) Heart with (r, p)-norm. (e) Sonar with ℓ_p-norm. (f) Sonar with (r, p)-norm. (g) Pima with ℓ_p-norm. (h) Pima with (r, p)-norm. (i) German with ℓ_p-norm. (j) German with (r, p)-norm. (k) Liver with ℓ_p-norm. (l) Liver with (r, p)-norm. (m) wdbc with ℓ_p-norm. (n) wdbc with (r, p)-norm.

out of seven data sets and inferior to (r, p)-MKL1 on one out of seven data sets, where the one in the middle represents that the two methods achieve the same performances on one out of the seven data sets. When compared to (r, p)-MKL2, (r, p)-MKL3 obtains 3-3-1, namely (r, p)-MKL3 outperforms (r, p)-MKL2 on three out of seven data sets, while inferior to (r, p)-MKL2 on one out of seven data sets, and achieve the same performances on three out of the seven data sets. Nevertheless, further analysis using t-test and Wilcoxons signed rank test shows that these differences are not statistically significant (namely Tie). The same holds true also for training time and support vector percentages, as shown in Table II(b) and (c), demonstrating that (11) and (12) are equivalent, and the proposed (r, p)-MKL is globally convergent and thus insensitive to the initialization of the kernel weights.

Note that the reported training time for (r, p)-MKL2 and (r, p)-MKL3 also includes the training time for the involved ℓ_p-MKL. But interestingly, this does not increase the whole training time significantly, which could be attributed to the faster convergence rate with ℓ_p-MKL initialization.

B. Comparison With Vector ℓ_p-MKL

We also compare (r, p)-MKL with the optimal implementation strategy against canonical ℓ_p-MKL [33]. For fair comparison, the regularization parameter in ℓ_p-MKL is jointly optimized using cross validation over: $C \in \{0.01, 0.1, 1, 10, 100\}$ and $p \in \{1, 2, 4, 10\}$ as well. Table III lists the optimized p and (r, p) values in ℓ_p-MKL and (r, p)-MKL, respectively, for each of the seven UCI data sets, from which we can see that these values are data-dependent, and thus to a certain extent reveal the intrinsic sparsity and the order of the kernel pair relationship of the data. As mentioned in Section II, (r, p)-MKL will be degenerated into canonical ℓ_p-MKL when $r = p$, and thus $(1, 1)$-MKL is the sparse case of (r, p)-MKL. Hence, Table III just indicates that under both vector and matrix regularization nonsparse MKL performs better than its sparse counterpart.

Different performance measures, including average testing accuracies (ACC), training time (Time), support vector percentages (SV), and the number of iterations required for convergence (# of Iter) for multiple linear kernels (K_L), multiple polynomial kernels (K_P), and multiple Gaussian kernels (K_G) for each of the seven UCI data sets are reported in Table IV. In terms of testing accuracy, (r, p)-MKL shows consistent out-performances to ℓ_p-MKL for all the three types of multiple kernels involved. Taking multiple K_G values as an example, (r, p)-MKL is superior to ℓ_p-MKL on six out of seven data sets, while inferior to ℓ_p-MKL only on one out of seven data sets (namely $6-0-1$) using direct comparison. The statistical t-test further validates the statistically significant difference by a statistical test for each one of the seven classification tasks, that is (r, p)-MKL outperforms ℓ_p-MKL on three out of seven data sets, while achieving the same performances on four out of the seven data sets (namely $3-4-0$). The significance is further verified by Wilcoxon signed rank test, namely, Win. With multiple Gaussian kernels, Fig. 1 compares the weights distribution learned by ℓ_p-MKL and by (r, p)-MKL, respectively. Intuitively, there are significant differences between the two regularization strategies. Comparing (5) and (6), we can tell that such differences are caused mathematically by the cross terms, namely, the interaction of kernels. Hence, Fig. 1 provides us a qualitative illustration on how the interaction of kernels affects the learning results, and the quantitative performance gains obtained by such interaction are: 1.97% for Ionosphere, 0.74% for Heart, 0.09% for Pima, 4.17% for German, 2.76% for Liver disorder, and 0.79% for wdbc, respectively. The improvements of (r, p)-MKL to canonical vector ℓ_p-MKL suggest that the 2-D (r, p)-norm constraint boosts the overall accuracy.

In terms of training time and support vector percentages, there is no significant difference between (r, p)-MKL and ℓ_p-MKL, justifying the proposed (r, p)-MKL is as efficient as ℓ_p-MKL in both training and test. As shown in Table IV, only one to two outer iterations are required for convergence. This could be attributed to the use of closed-form solution to the kernel weights $\boldsymbol{\beta}$ and the convergence ensured by the convex formulation. For the matrix-regularized **Q**-MKL [29], a domain specific matrix **Q** has to be elaborately designed for each task. We have not found an empirical mechanism for fair comparison in diverse classification tasks.

VI. Conclusion

We presented a matrix regularized MKL via (r, p)-norms. This extends vector ℓ_p-norm regularization and helps explore the dependences and interactions among kernels leading to better performance. We gave a simple alternating optimization with closed-form solution for the kernel weights and shown the global convergence of the proposed problem that can always be guaranteed. We analyzed such a regularizer using a Rademacher complexity bound, and we also proved that (r, p)-norm MKL yields strictly better generalization bounds than ℓ_p-norm MKL. Finally, we reported the results of (r, p)-MKL on several publicly available data sets. (r, p)-MKL was shown to achieve consistently superior performances to canonical ℓ_p-MKL, demonstrating the benefits of revealing the higher order kernel-pair relationships. Nevertheless, this paper constitutes only a preliminary study and that a deeper analysis with more expressive formulation and efficient solving strategy should be further investigated.

Appendix

This section is based on the derivation of a generalized MKL approach in the dual space presented in [33]. Specifically, we first rewrite optimization problem (7) as follows:

$$\begin{aligned}\min_{\mathbf{w},\mathbf{b},\mathbf{t},\boldsymbol{\beta}} \quad & \sum_{i=1}^{N} \ell(t_i, y_i) + \frac{1}{2}\sum_{m=1}^{M} \frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m} \\ \text{s.t. } \forall i: & \sum_{m=1}^{M} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b = t_i \\ & \|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} \le 1, \quad \boldsymbol{\beta} \ge 0 \end{aligned} \tag{42}$$

where ℓ is a convex loss function.

The Lagrangian of problem (42) is

$$\begin{aligned}\mathcal{L} = {} & C\sum_{i=1}^{N} \ell(t_i, y_i) + \frac{1}{2}\sum_{m=1}^{M} \frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m} \\ & - \sum_{i=1}^{N} \alpha_i \left(\sum_{m=1}^{M} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b - t_i \right) \\ & + \lambda \left(\frac{1}{2}\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} - \frac{1}{2} \right) - \boldsymbol{\gamma}^T \boldsymbol{\beta} \end{aligned} \tag{43}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^N$, $\lambda \in \mathbb{R}_+$, and $\boldsymbol{\gamma} \in \mathbb{R}^M$ are the Lagrange multipliers of the constraints. Then setting to zero, the first partial derivatives of the Lagrangian $\mathcal{L}$ with respect to the primal variables $\mathbf{w}$ and b we have the following optimality conditions:

$$\begin{aligned} \mathbf{1}^T \boldsymbol{\alpha} &= 0 \\ \mathbf{w}_m &= \beta_m \sum_{i=1}^{N} \alpha_i \phi_m(\mathbf{x}_i) \quad \forall m = 1, \ldots, M. \end{aligned} \tag{44}$$

Resubstituting the above equations in the Lagrangian problem yields

$$\begin{aligned} \max_{\substack{\boldsymbol{\alpha},\lambda,\boldsymbol{\gamma}:\mathbf{1}^T\boldsymbol{\alpha}=0,\\ \lambda\ge0,\boldsymbol{\gamma}\ge0}} \min_{\mathbf{t},\boldsymbol{\beta}} \; & C\sum_{i=1}^{N} (\ell(t_i, y_i) + \alpha_i t_i) - \frac{1}{2}\sum_{m=1}^{M} \beta_m \boldsymbol{\alpha}^T \mathbf{K}_m \boldsymbol{\alpha} \\ & + \lambda \left(\frac{1}{2}\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} - \frac{1}{2} \right) - \boldsymbol{\gamma}^T \boldsymbol{\beta}. \end{aligned} \tag{45}$$

Rewriting above problem we have

$$\begin{aligned} \max_{\substack{\boldsymbol{\alpha},\lambda,\boldsymbol{\gamma}:\mathbf{1}^T\boldsymbol{\alpha}=0,\\ \lambda\ge0,\boldsymbol{\gamma}\ge0}} \; & - C\sum_{i=1}^{N} \max_{t_i} \left(-\frac{\alpha_i}{C} t_i - \ell(t_i, y_i) \right) \\ & - \lambda \max_{\boldsymbol{\beta}} \left(\frac{1}{\lambda} \sum_{i=1}^{M} \left(\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{K}_m \boldsymbol{\alpha} + \gamma_m \right) \beta_m - \frac{1}{2}\|\boldsymbol{\beta}\boldsymbol{\beta}^T\|_{r,p} \right) - \frac{1}{2}\lambda. \end{aligned} \tag{46}$$

Since $h^*(\mathbf{x}) = \max_{\mathbf{u}} \mathbf{x}^T\mathbf{u} - h(\mathbf{u})$ is defined as the Fenchel–Legendre conjugate of a function h, we can express the above

Lagrangian as

$$\max_{\substack{\boldsymbol{\alpha},\lambda,\boldsymbol{\gamma}:\mathbf{1}^T\boldsymbol{\alpha}=0,\\ \lambda\geq 0,\boldsymbol{\gamma}\geq 0}} \; -C\sum_{i=1}^{N}\ell^*\left(-\frac{\alpha_i}{C}, y_i\right) - \frac{1}{\lambda}\left\|\mathbb{G}\mathbb{G}^T\right\|^*_{r,p} - \frac{1}{2}\lambda \tag{47}$$

where

$$\mathbb{G} = \left(\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}_m\boldsymbol{\alpha} + \gamma_m\right)_{m=1}^{M} \in \mathbb{R}^M$$

$\ell^*(\cdot)$ denotes the dual loss and $\|\cdot\|^*$ denotes the dual norm. By setting to zero the partial derivative of this Lagrangian with respect to λ, we get

$$\lambda = \left(\|\mathbb{G}\mathbb{G}^T\|^*_{r,p}\right)^{\frac{1}{2}} \tag{48}$$

from which we have the following generalized dual problem of (r, p)-MKL:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}:\mathbf{1}^T\boldsymbol{\alpha}=0,\boldsymbol{\gamma}\geq 0} -C\sum_{i=1}^{N}\ell^*\left(-\frac{\alpha_i}{C}, y_i\right) - \left(\|\mathbb{G}\mathbb{G}^T\|^*_{r,p}\right)^{\frac{1}{2}} \quad . \tag{49}$$

Furthermore, $\ell(t, y) = \max(0, 1 - ty)$ is the hinge loss and the dual loss of the hinge loss is [57]

$$\ell^*(t, y) = \begin{cases} \dfrac{t}{y} & \text{if } -1 \leq \frac{t}{y} \leq 0 \\ \infty & \text{elsewise.} \end{cases} \tag{50}$$

Then for each sample $(\mathbf{x}_i, y_i)$, the term $\ell^*(-(\alpha_i/C), y_i)$ in (49) translates to $-(\alpha_i/Cy_i)$ with $0 \leq (\alpha_i/y_i) \leq C$. Let $\alpha_i^{\text{new}} = (\alpha_i/y_i)$, and plugging this new variable in (49) gives

$$\begin{aligned} &\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}:\boldsymbol{\gamma}\geq 0} \; \mathbf{1}^T\boldsymbol{\alpha} - \left(\left\|\mathbb{G}\mathbb{G}^T\right\|^*_{r,p}\right)^{\frac{1}{2}} \\ &\text{s.t. } \mathbf{y}^T\boldsymbol{\alpha} = 0 \text{ and } \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1} \end{aligned} \tag{51}$$

where

$$\mathbb{G} = \left(\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^T\beta_m\mathbf{K}_m(\boldsymbol{\alpha}\circ\mathbf{y}) + \gamma_m\right)_{m=1}^{M} \in \mathbb{R}^M. \tag{52}$$

Finally, using the fact that the dual of (r, p)-norm is the (s, q)-norm, where s and q are the exponents dual to r and p, respectively, i.e., $(1/r) + (1/s) = 1$ and $(1/p) + (1/q) = 1$ [50], and note that $\boldsymbol{\gamma}^* = 0$ in the optimal point, we have the precisely dual form of (r, p)-MKL

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \; \mathbf{1}^T\boldsymbol{\alpha} - \|\mathbb{G}\mathbb{G}^T\|^{\frac{1}{2}}_{s,q} \\ &\text{s.t. } \mathbf{y}^T\boldsymbol{\alpha} = 0 \text{ and } \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1} \end{aligned} \tag{53}$$

where

$$\mathbb{G} = \left(\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^T\beta_m\mathbf{K}_m(\boldsymbol{\alpha}\circ\mathbf{y})\right)_{m=1}^{M} \in \mathbb{R}^M. \tag{54}$$

References

[1] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Jan. 2004.

[2] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, Jul. 2006.

[3] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[4] M. Hu, Y. Chen, and J. T. Y. Kwok, "Building sparse multiple-kernel SVM classifiers," *IEEE Trans. Neural Netw.*, vol. 20, no. 5, pp. 827–839, May 2009.

[5] H. Yang, Z. Xu, J. Ye, I. King, and M. R. Lyu, "Efficient sparse generalized multiple kernel learning," *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 433–446, Mar. 2011.

[6] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "A unifying framework for typical multitask multiple kernel learning problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1287–1297, Jul. 2014.

[7] X. Xu, I. W. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749–761, May 2013.

[8] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "Pareto-path multitask multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 51–61, Jan. 2015.

[9] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.

[10] Z. Xu, R. Jin, H. Yang, I. King, and M. R. Lyu, "Simple and efficient multiple kernel learning by group lasso," in *Proc. ICML*, 2010, pp. 1175–1182.

[11] C. Cortes, M. Mohri, and A. Rostamizadeh, "Two-stage learning kernel algorithms," in *Proc. ICML*, 2010, pp. 239–246.

[12] S. V. N. Vishwanathan, Z. Sun, N. Ampornpunt, and M. Varma, "Multiple kernel learning and the SMO algorithm," in *Proc. NIPS*, 2010, pp. 2361–2369.

[13] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. ICML*, 2009, p. 134.

[14] C. Cortes, M. Mohri, and A. Rostamizadeh, "Learning non-linear combinations of kernels," in *Proc. NIPS*, 2009, pp. 396–404.

[15] Y. Han and G. Liu, "Efficient learning of sample-specific discriminative features for scene classification," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 683–686, Nov. 2011.

[16] Y. Han and G. Liu, "Probability-confidence-kernel-based localized multiple kernel learning with l_p norm," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 3, pp. 827–837, Jun. 2012.

[17] Y. Han, K. Yang, and G. Liu, "L_p norm localized multiple kernel learning via semi-definite programming," *IEEE Signal Process. Lett.*, vol. 19, no. 10, pp. 688–691, Oct. 2012.

[18] Y. Han, K. Yang, Y. Ma, and G. Liu, "Localized multiple kernel learning via sample-wise alternating optimization," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 137–148, Jan. 2014.

[19] Y. Lei, A. Binder, Ü. Dogan, and M. Kloft, "Theory and algorithms for the localized setting of learning kernels," in *Proc. FE NIPS*, 2015, pp. 173–195.

[20] Y. Gu, T. Liu, X. Jia, J. A. Benediktsson, and J. Chanussot, "Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3235–3247, Jun. 2016.

[21] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proc. ICML*, 2004, p. 6.

[22] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, "Efficient and accurate l_p-norm multiple kernel learning," in *Proc. NIPS*, 2009, pp. 997–1005.

[23] Q. Wang, Y. Gu, and D. Tuia, "Discriminative multiple kernel learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3912–3927, Jul. 2016.

[24] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Proc. NIPS*, 2001, pp. 367–373.

[25] J. Ye, J. Chen, and S. Ji, "Discriminant kernel and regularization parameter learning via semidefinite programming," in *Proc. ICML*, 2007, pp. 1095–1102.

[26] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler, "L_p norm multiple kernel fisher discriminant analysis for object and image categorisation," in *Proc. CVPR*, Jun. 2010, pp. 3626–3632.

[27] Y. Gu, Q. Wang, H. Wang, D. You, and Y. Zhang, "Multiple kernel learning via low-rank nonnegative matrix factorization for classification of hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 8, no. 6, pp. 2739–2751, Jun. 2015.

[28] M. Kloft, U. Rückert, and P. L. Bartlett, "A unifying view of multiple kernel learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 953–997.

[29] C. Hinrichs, V. Singh, J. Peng, and S. C. Johnson, "Q-MKL: Matrix-induced regularization in multi-kernel learning with applications to neuroimaging," in *Proc. NIPS*, 2012, pp. 1430–1438.

[30] N. Srebro and S. Ben-David, "Learning bounds for support vector machines with learned kernels," in *Proc. COLT*, 2006, pp. 169–183.

[31] C. Cortes, M. Mohri, and A. Rostamizadeh, "L_2 regularization for learning kernels," in *Proc. UAI*, 2009, pp. 109–116.

[32] C. Cortes, M. Mohri, and A. Rostamizadeh, "Generalization bounds for learning kernels," in *Proc. ICML*, 2010, pp. 247–254.

[33] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "L_p-norm multiple kernel learning," *J. Mach. Learn. Res.*, vol. 12, pp. 953–997, Mar. 2011.

[34] Y. Han, K. Yang, Y. Yang, and Y. Ma, "Localized multiple kernel learning with dynamical clustering and matrix regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2635151.

[35] X. Li, G. Cui, and Y. Dong, "Graph regularized non-negative low-rank matrix factorization for image clustering," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3840–3853, May 2017.

[36] R. Zhang, F. Nie, and X. Li, "Regularized class-specific subspace classifier," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2738–2747, Nov. 2017.

[37] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy, "Composite kernel learning," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 73–103, 2010.

[38] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C. F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 563–574, Jun. 2012.

[39] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, "Regularization techniques for learning with matrices," *J. Mach. Learn. Res.*, vol. 13, pp. 1865–1890, Jun. 2012.

[40] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 201–233, 2002.

[41] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, "Linear algorithms for online multitask classification," *J. Mach. Learn. Res.*, vol. 11, pp. 2901–2934, Oct. 2010.

[42] A. Agarwal, A. Rakhlin, and P. Bartlett, "Matrix regularization techniques for online multitask learning," EECS Dept., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2008-138, Oct. 2008.

[43] M. K. Warmuth and D. Kuzmin, "Online variance minimization," *Mach. Learn.*, vol. 87, no. 1, pp. 1–32, 2012.

[44] X. Zhang, Y.-L. Yu, and D. Schuurmans, "Accelerated training for matrix-norm regularization: A boosting approach," in *Proc. NIPS*, 2012, pp. 2915–2923.

[45] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[46] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, Jun. 2008.

[47] Y. Han, K. Yang, Y. Yang, and Y. Ma, "On the impact of regularization variation on localized multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2688365.

[48] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectral–spatial classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2077–2081, Nov. 2017.

[49] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, Melbourne, VIC, Australia, Aug. 2017, pp. 2201–2207.

[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, Mar. 2004.

[51] M. Kloft, U. Rückert, and P. L. Bartlett, "A unifying view of multiple kernel learning," in *Proc. Eur. Conf. Mach. Learn. (ECML)*, Sep. 2010, pp. 66–81.

[52] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *Ann. Stat.*, vol. 30, no. 1, pp. 1–50, 2002.

[53] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

[54] M. Kloft and G. Blanchard, "On the convergence rate of l_p-norm multiple kernel learning," *J. Mach. Learn. Res.*, vol. 13, pp. 2465–2502, Aug. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2503321

[55] J. M. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. New York, NY, USA: Cambridge Univ. Press, 2004.

[56] M. Gönen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. ICML*, 2008, pp. 352–359.

[57] R. M. Rifkin and R. A. Lippert, "Value regularization and Fenchel duality," *J. Mach. Learn. Res.*, vol. 8, pp. 441–479, Mar. 2007.

Yina Han (M'16) received the B.S. and Ph.D. degrees in electronic and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2012, respectively.

From 2007 to 2008, she was a joint-training Ph.D. student with the Laboratoire Traitement et Communication de l'Information, Telecom Paris-Tech, Paris, France. She has been with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, since 2012, where she is currently an Associate Professor. Her current research interests include machine learning, pattern analysis, and signal and information processing.

Yixin Yang (M'03) received the B.S. degree in applied electronic engineering and the M.S. and Ph.D. degrees in underwater acoustic engineering from Northwestern Polytechnical University (NPU), Xi'an, China, in 1997, 1999, and 2002, respectively.

From 2002 to 2004, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has been with the School of Marine Science and Technology, NPU, since 2004, where he is currently a Professor. His current research interests include acoustic array signal processing, spectral estimation, and their applications.

Xuelong Li (M'02–SM'07–F'12) is currently a Full Professor with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China.

Qingyu Liu is currently the Director and a Senior Fellow with the Institute of Navy Research of China, Beijing, China. He is also the Distinguished Professor with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China.

Yuanliang Ma received the bachelor's degree in underwater acoustics from a joint educational program run by Northwestern Polytechnical University (NPU), Xi'an, China, and the former Harbin Institute of Military Engineering, Harbin, China, in 1961.

Since 1961, he has been involved in underwater acoustics and signal processing. From 1981 to 1983, he was a Visiting Scholar with the Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, U.K. In 1980, he became an Associate Professor with NPU, where he has been a Full Professor since 1985. His has authored three books, i.e., *Underwater Acoustic Transducers*, *Sensor Array Beam-Pattern Optimization: Theory with Applications*, and *Adaptive Active Noise Control*, in addition to over 300 journal and conference papers. His current research interests include the sensor array signal processing, ocean acoustics, microwave propagation, and signal processing systems.

Prof. Ma has been elected as a member of the Chinese Academy of Engineering. He is also a fellow of the Acoustical Society of China and a member of the Acoustical Society of America. He was a Vice-President of the Acoustical Society of China and the Chairman of its Underwater Acoustics Chapter from 1998 to 2006. He is currently the Chairman of the Academic Committee. He is an Associate Editor of the *Chinese Science Bulletin*.