

A Unifying View of Multiple Kernel Learning

Marius Kloft*, Ulrich Rückert, and Peter L. Bartlett

University of California, Berkeley, USA
{mkloft,rueckert,bartlett}@cs.berkeley.edu

Abstract. Recent research on multiple kernel learning has led to a number of approaches for combining kernels in regularized risk minimization. The proposed approaches include different formulations of objectives and varying regularization strategies. In this paper we present a unifying optimization criterion for multiple kernel learning and show how existing formulations are subsumed as special cases. We also derive the criterion’s dual representation, which is suitable for general smooth optimization algorithms. Finally, we evaluate multiple kernel learning in this framework analytically using a Rademacher complexity bound on the generalization error and empirically in a set of experiments.

1 Introduction

Selecting a suitable kernel for a kernel-based [17] machine learning task can be a difficult task. From a statistical point of view, the problem of choosing a good kernel is a model selection task. To this end, recent research has come up with a number of *multiple kernel learning* (MKL) [11] approaches, which allow for an automated selection of kernels from a predefined family of potential candidates. Typically, MKL approaches come in one of these three different flavors:

- (I) Instead of formulating an optimization criterion with a fixed kernel k , one leaves the choice of k as a variable and demands that k is taken from a linear span of base kernels $k := \sum_{i=1}^M \theta_i k_i$. The actual learning procedure then optimizes not only over the parameters of the kernel classifier, but also over θ subject to the constraint that $\|\theta\| \leq 1$ for some fixed norm. This approach is taken in [14, 20] for 1-norm penalties and extended in [9] to ℓ_p -norms.
- (II) A second approach takes k from a (non-)linear span of base kernels $k := \sum_{i=1}^M \theta_i^{-1} k_i$ subject to the constraint that $\|\theta\| \leq 1$ for some fixed norm. This approach was taken in [2] and [13] for p -norms and ℓ_∞ norm, respectively.
- III) A third approach optimizes over all kernel classifiers for each of the M base kernels, but modifies the regularizer to a block norm, that is, a norm of the vector containing the individual kernel norms. This allows to trade-off the contributions of each kernel to the final classifier. This formulation was used, for example, in [4].

* Also at Machine Learning Group, Technische Universität Berlin, Franklinstr. 28/29, FR 6-9, 10587 Berlin, Germany.

- (IV) Finally, since it appears to be sensible to have only the best kernels contribute to the final classifier, it makes sense to encourage sparse kernel weights. One way to do so is to extend the second setting with an *elastic net* regularizer, a linear combination of ℓ_1 and ℓ_2 regularizers. This approach was first considered in [4] as a numerical tool to approximate the ℓ_1 -norm constraint and subsequently analyzed in [22] for its regularization properties.

While all of these formulations are based on similar considerations, the individual formulations and used techniques vary considerably. The particular formulations are tailored more towards a specific optimization approach rather than the inherent characteristics. Type (I) and (II) approaches, for instance, are generally solved using partially dualized wrapper approaches; (III) is directly optimized the dual; and (IV) solves MKL in the primal, extending the approach of [6]. This makes it hard to gain insights into the underpinnings and differences of the individual methods, to design general-purpose optimization procedures for the various criteria and to compare the different techniques empirically.

In this paper, we show that all the above approaches can be viewed under a common umbrella by extending the block norm framework (III) to more general norms; we thus formulate MKL as an optimization criterion with a block-norm regularizer. By using this specific form of regularization, we can incorporate all the previously mentioned formulations as special cases of a single criterion. We derive a modular dual representation of the criterion, which separates the contribution of the loss function and the regularizer. This allows practitioners to plug in specific (dual) loss functions and to adjust the regularizer in a flexible fashion.

We show how the dual optimization problem can be solved using standard smooth optimization techniques, report on experiments on real world data, and compare the various approaches according to their ability to recover sparse kernel weights. On the theoretical side, we give a concentration inequality that bounds the generalization ability of MKL classifiers obtained in the presented framework. The bound is the first known bound to apply to MKL with elastic net regularization; it matches the best previously known bound [8] for the special case of ℓ_1 and ℓ_2 regularization, and it is the first bound for ℓ_p block norm MKL with arbitrary p .

2 Multiple Kernel Learning—A Unifying View

In this section we cast multiple kernel learning in a unified framework. Before we go into the details, we need to introduce the general setting and notation.

2.1 MKL in the Primal

We begin with reviewing the classical supervised learning setup. Given a labeled sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1, \dots, n}$, where the \mathbf{x}_i lie in some input space \mathcal{X} and $y_i \in$

$\mathcal{Y} \subset \mathbb{R}$, the goal is to find a hypothesis $f \in \mathcal{H}$, that generalizes well on new and unseen data. Regularized risk minimization returns a minimizer f^* ,

$$f^* \in \operatorname{argmin}_f \operatorname{R}_{\text{emp}}(f) + \lambda \Omega(f),$$

where $\operatorname{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i)$ is the empirical risk of hypothesis f w.r.t. a convex loss function $\ell : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$, $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is a regularizer, and $\lambda > 0$ is a trade-off parameter. We consider linear models of the form

$$f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle, \quad (1)$$

together with a (possibly non-linear) mapping $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ to a Hilbert space \mathcal{H} [18, 12] and constrain the regularization to be of the form $\Omega(f) = \frac{1}{2} \|\mathbf{w}\|_2^2$ which allows to kernelize the resulting models and algorithms. We will later make use of kernel functions $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$ to compute inner products in \mathcal{H} . When learning with multiple kernels, we are given M different feature mappings $\Phi_m : \mathcal{X} \rightarrow \mathcal{H}_m$, $m = 1, \dots, M$, each giving rise to a reproducing kernel k_m of \mathcal{H}_m .

There are two main ways to formulate regularized risk minimization with MKL. The first approach, denoted by (I) in the introduction, introduces a linear kernel mixture $k_{\boldsymbol{\theta}} = \sum_{m=1}^M \theta_m k_m$, $\theta_m \geq 0$ and a blockwise weighted target vector $\mathbf{w}_{\boldsymbol{\theta}} := (\sqrt{\theta_1} \mathbf{w}_1^\top, \dots, \sqrt{\theta_M} \mathbf{w}_M^\top)^\top$. With this, one solves

$$\begin{aligned} \inf_{\mathbf{w}, \boldsymbol{\theta}} \quad & C \sum_{i=1}^n \ell \left(\sum_{m=1}^M \langle \sqrt{\theta_m} \mathbf{w}_m, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}_m}, y_i \right) + \|\mathbf{w}_{\boldsymbol{\theta}}\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_q \leq 1. \end{aligned} \quad (2)$$

Alternatively, one can omit the explicit mixture vector $\boldsymbol{\theta}$ and use block-norm regularization instead (this approach was denoted by (III) in the introduction). In this case, denoting by $\|\mathbf{w}\|_{2,p} = \left(\sum_{m=1}^M \|\mathbf{w}_m\|_{\mathcal{H}_m}^p \right)^{1/p}$ the ℓ_2/ℓ_p block norm, one optimizes

$$\inf_{\mathbf{w}} \quad C \sum_{i=1}^n \ell \left(\sum_{m=1}^M \langle \mathbf{w}_m, \Phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m}, y_i \right) + \|\mathbf{w}\|_{2,p}^2. \quad (3)$$

One can show that (2) is a special case of (3). In particular, one can show that setting the block-norm parameter to $p = \frac{2q}{q+1}$ is equivalent to having kernel mixture regularization with $\|\boldsymbol{\theta}\|_q \leq 1$ [10]. This also implies that the kernel mixture formulation is *strictly* less general, because it can not replace block norm regularization for $p > 2$.

Hence, we focus on the block norm criterion, and extend it to also include elastic net regularization. The resulting primal problem generalizes the approaches (I)–(IV); it is stated as follows:

Primal MKL Optimization Problem

$$\inf_{\mathbf{w}} C \sum_{i=1}^n \ell(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}}, y_i) + \frac{1}{2} \|\mathbf{w}\|_{2,p}^2 + \frac{\mu}{2} \|\mathbf{w}\|_2^2, \quad (\text{P})$$

where $\Phi = \Phi_1 \times \cdots \times \Phi_M$ denotes the Cartesian product of the Φ_m 's. Using the above criterion it is possible to recover block norm regularization by setting $\mu = 0$ and the elastic net regularizer by setting $p = 1$.

Note that we use a slightly different—but equivalent—regularization than the one used in the original elastic net paper [25]: we square the term $\|\mathbf{w}\|_{2,p}$ while in the original criterion it appeared linearly. To see that the two formulations are equal, notice that the original regularizer can equivalently be encoded as a hard constraint $\|\mathbf{w}\|_{2,p} \leq \eta$ (this is similar to a well known result for SVMs; see [23]), which is equivalent to $\|\mathbf{w}\|_{2,p}^2 < \eta^2$ and subsequently can be incorporated into the objective, again. Hence, it is equivalent: regularizing with $\|\mathbf{w}\|_{2,p}$ and $\|\mathbf{w}\|_{2,p}^2$, respectively, leads to the same regularization path.

2.2 MKL in Dual Space

Optimization problems often have a considerably easier structure when studied in the dual space. In this section we derive the dual problem of the generalized MKL approach presented in the previous section. Let us begin with rewriting Optimization Problem (P) by expanding the decision values into slack variables as follows,

$$\begin{aligned} \inf_{\mathbf{w}, \mathbf{t}} C \sum_{i=1}^n \ell(t_i, y_i) + \frac{1}{2} \|\mathbf{w}\|_{2,p}^2 + \frac{\mu}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t. } \forall i : \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}} = t_i. \end{aligned} \quad (4)$$

Applying Lagrange's theorem re-incorporates the constraints into the objective by introducing Lagrangian multipliers $\alpha \in \mathbb{R}^n$. The Lagrangian saddle point problem is then given by

$$\begin{aligned} \sup_{\alpha} \inf_{\mathbf{w}, \mathbf{t}} C \sum_{i=1}^n \ell(t_i, y_i) + \frac{1}{2} \|\mathbf{w}\|_{2,p}^2 + \frac{\mu}{2} \|\mathbf{w}\|_2^2 \\ - \sum_{i=1}^n \alpha_i (\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}} - t_i). \end{aligned} \quad (5)$$

Setting the first partial derivatives of the above Lagrangian to zero w.r.t. \mathbf{w} gives the following KKT optimality condition

$$\forall m : \mathbf{w}_m = \left(\|\mathbf{w}\|_{2,p}^{2-p} \|\mathbf{w}_m\|^{p-2} + \mu \right)^{-1} \sum_i \alpha_i \Phi_m(\mathbf{x}_i). \quad (\text{KKT})$$

Inspecting the above equation reveals the representation $\mathbf{w}_m^* \in \text{span}(\Phi_m(\mathbf{x}_1), \dots, \Phi_m(\mathbf{x}_n))$. Rearranging the order of terms in the Lagrangian,

$$\begin{aligned} \sup_{\boldsymbol{\alpha}} \quad & -C \sum_{i=1}^n \sup_{\mathbf{t}} \left(-\frac{\alpha_i t_i}{C} - \ell(t_i, y_i) \right) \\ & - \sup_{\mathbf{w}} \left(\langle \mathbf{w}, \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}} - \frac{1}{2} \|\mathbf{w}\|_{2,p}^2 - \frac{\mu}{2} \|\mathbf{w}\|_2^2 \right), \end{aligned}$$

lets us express the Lagrangian in terms of Fenchel-Legendre conjugate functions $h^*(\mathbf{x}) = \sup_{\mathbf{u}} \mathbf{x}^\top \mathbf{u} - h(\mathbf{u})$ as follows,

$$\sup_{\boldsymbol{\alpha}} \quad -C \sum_{i=1}^n \ell^* \left(-\frac{\alpha_i}{C}, y_i \right) - \left(\frac{1}{2} \left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \right\|_{2,p}^2 + \frac{\mu}{2} \left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \right\|_2^2 \right)^*, \quad (6)$$

thereby removing the dependency of the Lagrangian on \mathbf{w} . The function ℓ^* is called *dual loss* in the following. Recall that the Inf-Convolution [16] of two functions f and g is defined by

$$(f \oplus g)(x) := \inf_y f(x - y) + g(y), \quad (7)$$

and that $(f^* \oplus g^*)(x) = (f + g)^*(x)$ and $(\eta f)^*(x) = \eta f^*(x/\eta)$ hold. Moreover, we have for the conjugate of the block norm $(\frac{1}{2} \|\cdot\|_{2,p}^2)^* = \frac{1}{2} \|\cdot\|_{2,p^*}^2$ [3] where p^* is the conjugate exponent, i.e., $\frac{1}{p} + \frac{1}{p^*} = 1$. As a consequence, we obtain the following *dual* optimization problem

Dual MKL Optimization Problem

$$\sup_{\boldsymbol{\alpha}} \quad -C \sum_{i=1}^n \ell^* \left(-\frac{\alpha_i}{C}, y_i \right) - \left(\frac{1}{2} \|\cdot\|_{2,p^*}^2 \oplus \frac{1}{2\mu} \|\cdot\|_2^2 \right) \left(\sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \right). \quad (D)$$

Note that the supremum is also a maximum, if the loss function is continuous. The function $f \oplus \frac{1}{2\mu} \|\cdot\|_2^2$ is the so-called *Moreau-Yosida Approximate* [19] and has been studied extensively both theoretically and algorithmically for its favorable regularization properties. It can “smoothen” an optimization problem—even if it is initially non-differentiable—and increase the condition number of the Hessian for twice differentiable problems.

The above dual generalizes multiple kernel learning to arbitrary convex loss functions and regularizers. Due to the mathematically clean separation of the loss and the regularization term—each loss term solely depends on a single real valued variable—we can immediately recover the corresponding dual for a specific choice of a loss/regularizer pair $(\ell, \|\cdot\|_{2,p})$ by computing the pair of conjugates $(\ell^*, \|\cdot\|_{2,p^*})$.

2.3 Obtaining Kernel Weights

While formalizing multiple kernel learning with block-norm regularization offers a number of conceptual and analytical advantages, it requires an additional step in practical applications. The reason for this is that the block-norm regularized dual optimization criterion does not include explicit kernel weights. Instead, this information is contained only implicitly in the optimal kernel classifier parameters, as output by the optimizer. This is a problem, for instance if one wishes to apply the induced classifier on new test instances. Here we need the kernel weights to form the final kernel used for the actual prediction. To recover the underlying kernel weights, one essentially needs to identify which kernel contributed to which degree for the selection of the optimal dual solution. Depending on the actual parameterization of the primal criterion, this can be done in various ways.

We start by reconsidering the KKT optimality condition given by Eq. (KKT) and observe that the first term on the right hand side,

$$\theta_m := \left(\|\mathbf{w}\|_{2,p}^{2-p} \|\mathbf{w}_m\|^{p-2} + \mu \right)^{-1}. \quad (8)$$

introduces a scaling of the feature maps. With this notation, it is easy to see from Eq. (KKT) that our model given by Eq. (1) extends to

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{m=1}^M \sum_{i=1}^n \alpha_i \theta_m k_m(\mathbf{x}_i, \mathbf{x}).$$

In order to express the above model solely in terms of dual variables we have to compute θ in terms of α .

In the following we focus on two cases. First, we consider ℓ_p block norm regularization for arbitrary $1 < p < \infty$ while switching the elastic net off by setting the parameter $\mu = 0$. Then, from Eq. (KKT) we obtain

$$\|\mathbf{w}_m\| = \|\mathbf{w}\|_{2,p}^{\frac{p-2}{p-1}} \left\| \sum_{i=1}^n \alpha_i \Phi_m(\mathbf{x}_i) \right\|_{\mathcal{H}_m}^{\frac{1}{p-1}} \quad \text{where} \quad \mathbf{w}_m = \theta_m \sum_i \alpha_i \Phi_m(\mathbf{x}_i).$$

Resubstitution into (8) leads to the proportionality

$$\exists c > 0 \quad \forall m : \quad \theta_m = c \left(\left\| \sum_{i=1}^n \alpha_i \Phi_m(\mathbf{x}_i) \right\|_{\mathcal{H}_m} \right)^{\frac{2-p}{p-1}}. \quad (9)$$

Note that, in the case of classification, we only need to compute θ up to a positive multiplicative constant.

For the second case, let us now consider the elastic net regularizer, i.e., $p = 1 + \epsilon$ with $\epsilon \approx 0$ and $\mu > 0$. Then, the optimality condition given by Eq. (KKT) translates to

$$\mathbf{w}_m = \theta_m \sum_i \alpha_i \Phi_m(\mathbf{x}_i) \quad \text{where} \quad \theta_m = \left(\left(\sum_{m'=1}^M \|\mathbf{w}_{m'}\|_{\mathcal{H}_{m'}}^{1+\epsilon} \right)^{1-\epsilon} \|\mathbf{w}_m\|_{\mathcal{H}_m}^{\epsilon-1} + \mu \right)^{-1}.$$

Inserting the left hand side expression for $\|\mathbf{w}_m\|_{\mathcal{H}_m}$ into the right hand side leads to the non-linear system of equalities

$$\forall m : \mu\theta_m\|K_m\|^{1-\epsilon} + \theta_m^\epsilon \left(\sum_{m'=1}^M \theta_{m'}^{1+\epsilon}\|K_{m'}\|^{1+\epsilon} \right)^{1-\epsilon} = \|K_m\|^{1-\epsilon}, \quad (10)$$

where we employ the notation $\|K_m\| := \|\sum_{i=1}^n \alpha_i \Phi_m(\mathbf{x}_i)\|_{\mathcal{H}_m}$. In our experiments we solve the above conditions numerically using $\epsilon \approx 0$. Notice, that this difficulty does not arise in [4] for $p = 1$ and in [22], which is an advantage of the latter approaches. The optimal mixing coefficients θ_m can now be computed solely from the dual $\boldsymbol{\alpha}$ variables by means of Eq. (9) and (10), and by the kernel matrices K_m using the identity

$$\forall m = 1, \dots, M : \|K_m\| = \sqrt{\boldsymbol{\alpha} K_m \boldsymbol{\alpha}}.$$

This enables optimization in the dual space as discussed in the next section.

3 Optimization Strategies

In this section we describe how one can simply solve the dual optimization problem by a common purpose quasi-Newton method. We do not claim that this is the fastest possible way to solve the problem; in the contrary, we conjecture that a SMO-type algorithm decomposition algorithm, as used in [4], might speed up the optimization. However, computational efficiency is not the focus of this paper; we focus on understanding and theoretically analyzing MKL and leave a more efficient implementation of our approach to future work.

For our experiments, we use the hinge loss $l(x) = \max(0, 1 - x)$ to obtain a support vector formulation, but the discussion also applies to most other convex loss functions. We first note that the dual loss of the hinge loss is $\ell^*(t, y) = \frac{t}{y}$ if $-1 \leq \frac{t}{y} \leq 0$ and ∞ otherwise [15]. Hence, for each i the term $\ell^*\left(-\frac{\alpha_i}{C}, y_i\right)$ of the generalized dual, i.e., Optimization Problem (D), translates to $-\frac{\alpha_i}{C y_i}$, provided that $0 \leq \frac{\alpha_i}{y_i} \leq C$. Employing a variable substitution of the form $\alpha_i^{\text{new}} = \frac{\alpha_i}{y_i}$, the dual problem (D) becomes

$$\sup_{\boldsymbol{\alpha}: \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}} \mathbf{1}^\top \boldsymbol{\alpha} - \left(\frac{1}{2} \|\cdot\|_{2,p^*}^2 \oplus \frac{1}{2\mu} \|\cdot\|_2^2 \right) \left(\sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{x}_i) \right),$$

and by definition of the Inf-convolution,

$$\sup_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}} \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2} \left\| \sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{x}_i) - \boldsymbol{\beta} \right\|_{2,p^*}^2 - \frac{1}{2\mu} \|\boldsymbol{\beta}\|_2^2. \quad (11)$$

We note that the representer theorem [17] is valid for the above problem, and hence the solution of (11) can be expressed in terms of kernel functions, i.e.,

$\beta_m = \sum_{i=1}^n \gamma_i k_m(x_i, \cdot)$ for certain real coefficients $\gamma \in \mathbb{R}^n$ uniformly for all m , hence $\beta = \sum_{i=1}^n \gamma_i \Phi(x_i)$. Thus, Eq. (11) has a representation of the form

$$\sup_{\alpha, \gamma: \mathbf{0} \leq \alpha \leq C \mathbf{1}} \mathbf{1}^\top \alpha - \frac{1}{2} \left\| \sum_{i=1}^n (\alpha_i y_i - \gamma_i) \Phi(x_i) \right\|_{2, p^*}^2 - \frac{1}{2\mu} \gamma^\top K \gamma,$$

where we use the shorthand $K = \sum_{m=1}^M K_m$. The above expression can be written¹ in terms of kernel matrices as follows,

Support Vector MKL—The Hinge Loss Dual

$$\sup_{\alpha, \gamma: \mathbf{0} \leq \alpha \leq C \mathbf{1}} \mathbf{1}^\top \alpha - \frac{1}{2} \left\| ((\alpha \circ \mathbf{y} - \gamma)^\top K_m (\alpha \circ \mathbf{y} - \gamma))_{m=1}^M \right\|_{\ell_2^*} - \frac{1}{2\mu} \gamma^\top K \gamma. \quad (\text{SV-MKL})$$

In our experiments, we optimized the above criterion by using the limited memory quasi-Newton software L-BFGS-B [24]. L-BFGS-B is a common purpose solver that can simply be used out-of-the-box. It approximates the Hessian matrix based on the last t gradients, where t is a parameter to be chosen by the user. Note that L-BFGS-B can handle the box constraint induced by the hinge loss.

4 Theoretical Analysis

In this section we give two uniform convergence bounds for the generalization error of the multiple kernel learning formulation presented in Section 2. The results are based on the established theory on Rademacher complexities. Let $\sigma_1, \dots, \sigma_n$ be a set of independent Rademacher variables, which obtain the values -1 or +1 with the same probability 0.5, and let \mathcal{C} be some space of classifiers $c: \mathcal{X} \rightarrow \mathbb{R}$. Then, the *Rademacher complexity* of \mathcal{C} is given by

$$\mathcal{R}_{\mathcal{C}} := \mathbf{E} \left[\sup_{c \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i c(x_i) \right].$$

If the Rademacher complexity of a class of classifiers is known, it can be used to bound the generalization error. We give one result here, which is an immediate corollary of Thm. 8 in [5] (using Thm. 12.4 in the same paper), and refer to the literature [5] for further results on Rademacher penalization.

Theorem 1. *Assume the loss $\ell: \mathbb{R} \supseteq \mathcal{Y} \rightarrow [0, 1]$ is Lipschitz with constant L . Then, the following holds with probability larger than $1 - \delta$ for all classifiers $c \in \mathcal{C}$:*

$$\mathbf{E}[\ell(y c(x))] \leq \frac{1}{n} \sum_{i=1}^n \ell(y_i c(x_i)) + 2L \mathcal{R}_{\mathcal{C}} + \sqrt{\frac{8 \ln \frac{2}{\delta}}{n}}. \quad (12)$$

¹ We employ the notation $s = (s_1, \dots, s_M)^\top = (s_m)_{m=1}^M$ for $s \in \mathbb{R}^M$ and denote by $\mathbf{x} \circ \mathbf{y}$ the elementwise multiplication of two vectors.

We will now give an upper bound for the Rademacher complexity of the block-norm regularized linear learning approach described above. More precisely, for $1 \leq i \leq M$ let $\|w\|_{\star i} := \sqrt{k_i(w, w)}$ denote the norm induced by kernel k_i and for $x \in \mathbb{R}^p$, $p, q \geq 1$ and $C_1, C_2 \geq 0$ with $C_1 + C_2 = 1$ define

$$\|x\|_O := C_1 \|x\|_p + C_2 \|x\|_q.$$

We now give a bound for the following class of linear classifiers:

$$\mathcal{C}_\star := \left\{ c : \begin{pmatrix} \Phi_1(x) \\ \vdots \\ \Phi_M(x) \end{pmatrix} \mapsto \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}^T \begin{pmatrix} \Phi_1(x) \\ \vdots \\ \Phi_M(x) \end{pmatrix} \left\| \begin{pmatrix} \|w_1\|_{\star 1} \\ \vdots \\ \|w_M\|_{\star M} \end{pmatrix} \right\|_O \leq 1 \right\}.$$

Theorem 2. *Assume the kernels are normalized, i.e. $k_i(x, x) = \|x\|_{\star i}^2 \leq 1$ for all $x \in \mathcal{X}$ and all $1 \leq i \leq M$. Then, the Rademacher complexity of the class \mathcal{C}_\star of linear classifiers with block norm regularization is upper-bounded as follows:*

$$\mathcal{R}_{\mathcal{C}_\star} \leq \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \left(\sqrt{\frac{2 \ln M}{n}} + \sqrt{\frac{1}{n}} \right). \quad (13)$$

For the special case with $p \geq 2$ and $q \geq 2$, the bound can be improved as follows:

$$\mathcal{R}_{\mathcal{C}_\star} \leq \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \sqrt{\frac{1}{n}}. \quad (14)$$

Interpretation of Bounds. It is instructive to compare this result to some of the existing MKL bounds in the literature. For instance, the main result in [8] bounds the Rademacher complexity of the ℓ_1 -norm regularizer with a $O(\sqrt{\ln M/n})$ term. We get the same result by setting $C_1 = 1, C_2 = 0$ and $p = 1$. For the ℓ_2 -norm regularized setting, we can set $C_1 = 1, C_2 = 0$ and $p = \frac{4}{3}$ (because the kernel weight formulation with ℓ_2 norm corresponds to the block-norm representation with $p = \frac{4}{3}$) to recover their $O(M^{\frac{1}{4}}/\sqrt{n})$ bound. Finally, it is interesting to see how changing the C_1 parameter influences the generalization capacity of the elastic net regularizer ($p = 1, q = 2$). For $C_1 = 1$, we essentially recover the ℓ_1 regularization penalty, but as C_1 approaches 0, the bound includes an additional $O(\sqrt{M})$ term. This shows how the capacity of the elastic net regularizer increases towards the ℓ_2 setting with decreasing sparsity.

Proof (of Theorem 2). Using the notation $w := (w_1, \dots, w_M)^T$ and $\|w\|_B := \|(\|w_1\|_{\star 1}, \dots, \|w_M\|_{\star M})^T\|_O$ it is easy to see that

$$\begin{aligned} \mathbf{E} \left[\sup_{c \in \mathcal{C}_\star} \frac{1}{n} \sum_{i=1}^n \sigma_i y_i c(x_i) \right] &= \mathbf{E} \left[\sup_{\|w\|_B \leq 1} \left\{ \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}^T \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_1(x_i) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_M(x_i) \end{pmatrix} \right\} \right] \\ &= \mathbf{E} \left[\left\| \begin{pmatrix} \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_1(x_i) \right\|_{\star 1} \\ \vdots \\ \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_M(x_i) \right\|_{\star M} \end{pmatrix} \right\|_O^* \right], \end{aligned}$$

where $\|x\|^* := \sup_z \{z^T x \mid \|z\| \leq 1\}$ denotes the dual norm of $\|\cdot\|$ and we use the fact that $\|w\|_B^* = \|(\|w_1\|_{\star 1}^*, \dots, \|w_M\|_{\star M}^*)^T\|_O^*$ [3], and that $\|\cdot\|_{\star i}^* = \|\cdot\|_{\star i}$. We will show that this quantity is upper bounded by

$$\frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \left(\sqrt{\frac{2 \ln M}{n}} + \sqrt{\frac{1}{n}} \right). \quad (15)$$

As a first step we prove that for any $x \in \mathbb{R}^M$

$$\|x\|_O^* \leq \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \|x\|_\infty. \quad (16)$$

For any $a \geq 1$ we can apply Hölder's inequality to the dot product of $x \in \mathbb{R}^M$ and $\mathbf{1}_M := (1, \dots, 1)^T$ and obtain $\|x\|_1 \leq \|\mathbf{1}_M\|_{\frac{a}{a-1}} \cdot \|x\|_a = M^{\frac{a-1}{a}} \|x\|_a$. Since $C_1 + C_2 = 1$, we can apply this twice on the two components of $\|\cdot\|_O$ to get a lower bound for $\|x\|_O$,

$$(C_1 M^{\frac{1-p}{p}} + C_2 M^{\frac{1-q}{q}}) \|x\|_1 \leq C_1 \|x\|_p + C_2 \|x\|_q = \|x\|_O.$$

In other words, for every $x \in \mathbb{R}^M$ with $\|x\|_O \leq 1$ it holds that

$$\|x\|_1 \leq 1 / \left(C_1 M^{\frac{1-p}{p}} + C_2 M^{\frac{1-q}{q}} \right) = M / \left(C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}} \right).$$

Thus,

$$\{z^T x \mid \|x\|_O \leq 1\} \subseteq \left\{ z^T x \mid \|x\|_1 \leq \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \right\}. \quad (17)$$

This means we can bound the dual norm $\|\cdot\|_O^*$ of $\|\cdot\|_O$ as follows:

$$\begin{aligned} \|x\|_O^* &= \sup_z \{z^T x \mid \|z\|_O \leq 1\} \\ &\leq \sup_z \left\{ z^T x \mid \|z\|_1 \leq \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \right\} \\ &= \frac{M}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \|x\|_\infty. \end{aligned} \quad (18)$$

This accounts for the first factor in (15). For the second factor, we show that

$$\mathbf{E} \left[\left\| \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_1(x_i) \|_{\star 1} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_M(x_i) \|_{\star M} \end{pmatrix} \right\|_\infty \right] \leq \sqrt{\frac{2 \ln M}{n}} + \sqrt{\frac{1}{n}}. \quad (19)$$

To do so, define

$$V_k := \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_k(x_i) \right\|_{\star k}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j k_k(x_i, x_j).$$

By the independence of the Rademacher variables it follows for all $k \leq M$,

$$\mathbf{E}[V_k] = \frac{1}{n^2} \sum_{i=1}^n \mathbf{E}[k_k(x_i, x_i)] \leq \frac{1}{n}. \quad (20)$$

In the next step we use a martingale argument to find an upper bound for $\sup_k [W_k]$ where $W_k := \sqrt{V_k} - \mathbf{E}[\sqrt{V_k}]$. For ease of notation, we write $\mathbf{E}_{(r)}[X]$ to denote the conditional expectation $\mathbf{E}[X|(x_1, \sigma_1), \dots, (x_r, \sigma_r)]$. We define the following martingale:

$$\begin{aligned} Z_k^{(r)} &:= \mathbf{E}_{(r)}[\sqrt{V_k}] - \mathbf{E}_{(r-1)}[\sqrt{V_k}] \\ &= \mathbf{E}_{(r)} \left[\left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_k(x_i) \right\|_{\star k} \right] - \mathbf{E}_{(r-1)} \left[\left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_k(x_i) \right\|_{\star k} \right]. \end{aligned} \quad (21)$$

The range of each random variable $Z_k^{(r)}$ is at most $\frac{2}{n}$. This is because switching the sign of σ_r changes only one summand in the sum from $-\Phi_k(x_r)$ to $+\Phi_k(x_r)$. Thus, the random variable changes by at most $\left\| \frac{2}{n} \Phi_k(x_r) \right\|_{\star k} \leq \frac{2}{n} k_k(x_r, x_r) \leq \frac{2}{n}$. Hence, we can apply Hoeffding's inequality, $\mathbf{E}_{(r-1)} \left[e^{s Z_k^{(r)}} \right] \leq e^{\frac{1}{2n^2} s^2}$. This allows us to bound the expectation of $\sup_k W_k$ as follows:

$$\begin{aligned} \mathbf{E}[\sup_k W_k] &= \mathbf{E} \left[\frac{1}{s} \ln \sup_k e^{s W_k} \right] \\ &\leq \mathbf{E} \left[\frac{1}{s} \ln \sum_{k=1}^M \exp \left[s \sum_{r=1}^n Z_k^{(r)} \right] \right] \\ &\leq \frac{1}{s} \ln \sum_{k=1}^M \prod_{r=1}^n \mathbf{E}_{(r)} \left[e^{s Z_k^{(r)}} \right] \\ &\leq \frac{1}{s} \ln \sum_{k=1}^M \left(e^{\frac{1}{2n^2} s^2} \right)^n \\ &= \frac{\ln M}{s} + \frac{s}{2n}, \end{aligned}$$

where we n times applied Hoeffding's inequality. Setting $s = \sqrt{2n \ln M}$ yields:

$$\mathbf{E}[\sup_k W_k] \leq \sqrt{\frac{2 \ln M}{n}}. \quad (22)$$

Now, we can combine (20) and (22):

$$\mathbf{E} \left[\sup_k \sqrt{V_k} \right] \leq \mathbf{E} \left[\sup_k W_k + \sqrt{\mathbf{E}[V_k]} \right] \leq \sqrt{\frac{2 \ln M}{n}} + \sqrt{\frac{1}{n}}.$$

This concludes the proof of (19) and therewith (13).

The special case (14) for $p, q \geq 2$ is similar. As a first step, we modify (16) to deal with the ℓ_2 -norm rather than the ℓ_∞ -norm:

$$\|x\|_O^* \leq \frac{\sqrt{M}}{C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}}} \|x\|_2. \quad (23)$$

To see this, observe that for any $x \in \mathbb{R}^M$ and any $a \geq 2$ Hölder's inequality gives $\|x\|_2 \leq M^{\frac{a-2}{2a}} \|x\|_a$. Applying this to the two components of $\|\cdot\|_O$ we have:

$$(C_1 M^{\frac{2-p}{2p}} + C_2 M^{\frac{2-q}{2q}}) \|x\|_2 \leq C_1 \|x\|_p + C_2 \|x\|_q = \|x\|_O.$$

In other words, for every $x \in \mathbb{R}^M$ with $\|x\|_O \leq 1$ it holds that

$$\|x\|_2 \leq 1 / \left(C_1 M^{\frac{2-p}{2p}} + C_2 M^{\frac{2-q}{2q}} \right) = \sqrt{M} / \left(C_1 M^{\frac{1}{p}} + C_2 M^{\frac{1}{q}} \right).$$

Following the same arguments as in (17) and (18) we obtain (23). To finish the proof it now suffices to show that

$$\mathbf{E} \left[\left\| \begin{pmatrix} \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_1(x_i) \right\|_{*1} \\ \vdots \\ \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_M(x_i) \right\|_{*M} \end{pmatrix} \right\|_2 \right] \leq \sqrt{\frac{M}{n}}.$$

This is can be seen by a straightforward application of (20):

$$\mathbf{E} \left[\sqrt{\sum_{k=1}^M \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_k(x_i) \right\|_{*k}^2} \right] \leq \sqrt{\mathbf{E} \left[\sum_{k=1}^M V_k \right]} \leq \sqrt{\sum_{i=1}^M \frac{1}{n}} = \sqrt{\frac{M}{n}}.$$

5 Empirical Results

In this section we evaluate the proposed method on artificial and real data sets. To avoid validating over two regularization parameters simultaneously, we only study elastic net MKL for the special case $p \approx 1$.

5.1 Experiments with Sparse and Non-Sparse Kernel Sets

The goal of this section is to study the relationship of the level of sparsity of the true underlying function to the chosen block norm or elastic net MKL model. Apart from investigating which parameter choice leads to optimal results, we are also interested in the effects of suboptimal choices of p . To this aim we constructed several artificial data sets in which we vary the degree of sparsity in the true kernel mixture coefficients. We go from having all weight focused on a single kernel (the highest level of sparsity) to uniform weights (the least sparse scenario possible) in several steps. We then study the statistical performance of ℓ_p -block-norm MKL for different values of p that cover the entire range $[0, \infty]$.

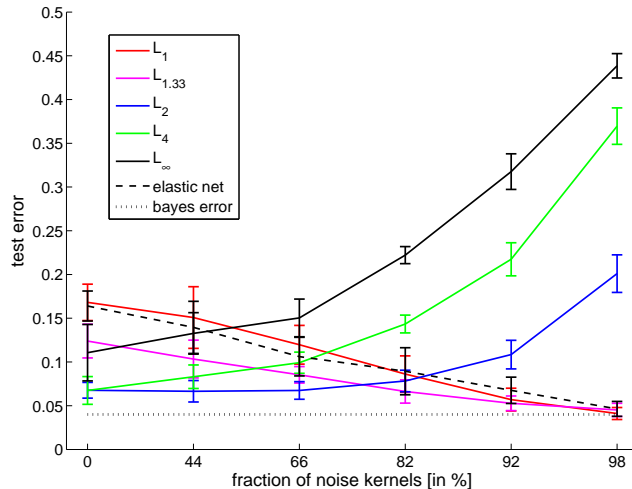


Fig. 1. Empirical results of the artificial experiment for varying true underlying data sparsity.

We follow the experimental setup of [10] but compute classification models for $p = 1, 4/3, 2, 4, \infty$ block-norm MKL and $\mu = 10$ elastic net MKL. The results are shown in Fig. 1 and compared to the Bayes error that is computed analytically from the underlying probability model.

Unsurprisingly, ℓ_1 performs best in the sparse scenario, where only a single kernel carries the whole discriminative information of the learning problem. In contrast, the ℓ_∞ -norm MKL performs best when all kernels are equally informative. Both MKL variants reach the Bayes error in their respective scenarios. The elastic net MKL performs comparable to ℓ_1 -block-norm MKL. The non-sparse $\ell_{4/3}$ -norm MKL and the unweighted-sum kernel SVM perform best in the balanced scenarios, i.e., when the noise level is ranging in the interval 60%-92%. The non-sparse ℓ_4 -norm MKL of [2] performs only well in the most non-sparse scenarios. Intuitively, the non-sparse $\ell_{4/3}$ -norm MKL of [7, 9] is the most robust MKL variant, achieving an test error of less than 0.1% in all scenarios. The sparse ℓ_1 -norm MKL performs worst when the noise level is less than 82%. It is worth mentioning that when considering the most challenging model/scenario combination, that is ℓ_∞ -norm in the sparse and ℓ_1 -norm in the uniformly non-sparse scenario, the ℓ_1 -norm MKL performs much more robust than its ℓ_∞ counterpart. However, as witnessed in the following sections, this does not prevent ℓ_∞ norm MKL from performing very well in practice. In summary, we conclude that by tuning the sparsity parameter p for each experiment, block norm MKL achieves a low test error across all scenarios.

5.2 Gene Start Recognition

This experiment aims at detecting transcription start sites (TSS) of RNA Polymerase II binding genes in genomic DNA sequences. Many detectors rely on a combination of feature sets which makes the learning task appealing for MKL. For our experiments we use the data set from [21] and we employ five different kernels representing the TSS signal (weighted degree with shift), the promoter (spectrum), the 1st exon (spectrum), angles (linear), and energies (linear). The kernel matrices are normalized such that each feature vector has unit norm in Hilbert space. We reserve 500 and 500 randomly drawn instances for holdout and test sets, respectively, and use 250 elemental training sets. Table 1 shows the area under the ROC curve (AUC) averaged over 250 repetitions of the experiment. Thereby 1 and ∞ block norms are approximated by 64/63 and 64 norms, respectively. For the elastic net we use an $\ell_{1.05}$ -block-norm penalty.

Table 1. Results for the bioinformatics experiment.

	AUC \pm stderr
$\mu = 0.01$ elastic net	85.91 \pm 0.09
$\mu = 0.1$ elastic net	85.77 \pm 0.10
$\mu = 1$ elastic net	87.73 \pm 0.11
$\mu = 10$ elastic net	88.24 \pm 0.10
$\mu = 100$ elastic net	87.57 \pm 0.09
1-block-norm MKL	85.77 \pm 0.10
4/3-block-norm MKL	87.93 \pm 0.10
2-block-norm MKL	87.57 \pm 0.10
4-block-norm MKL	86.33 \pm 0.10
∞ -block-norm MKL	87.67 \pm 0.09

The results vary greatly between the chosen MKL models. The elastic net model gives the best prediction for $\mu = 10$. Out of the block norm MKLs the classical ℓ_1 -norm MKL has the worst prediction accuracy and is even outperformed by an unweighted-sum kernel SVM (i.e., $p = 2$ norm MKL). In accordance with previous experiments in [9] the $p = 4/3$ -block-norm has the highest prediction accuracy of the models within the parameter range $p \in [1, 2]$. This performance can even be improved by the elastic net MKL with $\mu = 10$.

This is remarkable since elastic net MKL performs kernel *selection*, and hence the outputted kernel combination can be easily interpreted by domain experts. Note that the method using the unweighted sum of kernels [21] has recently been confirmed to be the leading in a comparison of 19 state-of-the-art promoter prediction programs [1]. It was recently shown to be outperformed by $\ell_{4/3}$ -norm MKL [9], and our experiments suggest that its accuracy can be further improved by $\mu = 10$ elastic net MKL.

6 Conclusion

We presented a framework for multiple kernel learning, that unifies several recent lines of research in that area. We phrased the seemingly different MKL variants as a single generalized optimization criterion and derived its dual representation. By plugging in an arbitrary convex loss function many existing approaches can be recovered as instantiations of our model. We compared the different MKL variants in terms of their generalization performance by giving an concentration inequality for generalized MKL that matches the previous known bounds for ℓ_1 and $\ell_{4/3}$ block norm MKL.

Our empirical analysis shows that the performance of the MKL variants crucially depends on true underlying data sparsity. We compared several existing MKL variants on bioinformatics data. On the computational side, we derived a quasi Newton optimization method for unified MKL. It is up to future work to speed up optimization by a SMO-type decomposition algorithm.

Acknowledgments

The authors wish to thank Francis Bach and Ryota Tomioka for comments that helped improving the manuscript; we thank Klaus-Robert Müller for stimulating discussions. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through the grant RU 1589/1-1 and by the European Community under the PASCAL2 Network of Excellence (ICT-216886). We gratefully acknowledge the support of NSF through grant DMS-0707060. MK acknowledges a scholarship by the German Academic Exchange Service (DAAD).

References

1. T. Abeel, Y. Van de Peer, and Y. Saeys. Towards a gold standard for promoter prediction evaluation. *Bioinformatics*, 2009.
2. J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. Saketha Nath, and S. Raman. Variable sparsity kernel learning — algorithms and applications. *Journal of Machine Learning Research*, 2010. Submitted. <http://mllab.csa.iisc.ernet.in/vskl.html>.
3. A. Agarwal, A. Rakhlin, and P. Bartlett. Matrix regularization techniques for online multitask learning. Technical Report UCB/EECS-2008-138, EECS Department, University of California, Berkeley, Oct 2008.
4. F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proc. 21st ICML*. ACM, 2004.
5. P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, November 2002.
6. O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 2006.
7. C. Cortes, M. Mohri, and A. Rostamizadeh. L2 regularization for learning kernels. In *Proceedings, 26th ICML*, 2009.

8. C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings, 27th ICML*, 2010. to appear. CoRR abs/0912.3309. <http://arxiv.org/abs/0912.3309>.
9. M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate lp-norm multiple kernel learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 997–1005. MIT Press, 2009.
10. M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Non-sparse regularization and efficient training with multiple kernels. Technical Report UCB/EECS-2010-21, EECS Department, University of California, Berkeley, Feb 2010. CoRR abs/1003.0079. <http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-21.html>.
11. G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
12. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 12(2):181–201, May 2001.
13. J. S. Nath, G. Dinesh, S. Ramanand, C. Bhattacharyya, A. Ben-Tal, and K. R. Ramakrishnan. On the algorithmics and applications of a mixed-norm based kernel learning formulation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 844–852, 2009.
14. A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.
15. R. M. Rifkin and R. A. Lippert. Value regularization and fenchel duality. *J. Mach. Learn. Res.*, 8:441–479, 2007.
16. R.T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, New Jersey, 1970.
17. B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
18. B. Schölkopf, A.J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
19. R. E. Showalter. Monotone operators in banach space and nonlinear partial differential equations. *Mathematical Surveys and Monographs*, 18, 1997.
20. S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7:1531–1565, July 2006.
21. S. Sonnenburg, A. Zien, and G. Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–e480, 2006.
22. R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in mkl. *arxiv*, 2010. CoRR abs/1001.2615.
23. V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
24. C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997.
25. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.