JMLR: Workshop and Conference Proceedings 60:1–16, 2016 ACML 2016

Localized Multiple Kernel Learning—A Convex Approach

Yunwen Lei YUNWELEI@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, China

Alexander Binder ALEXANDER_BINDER@SUTD.EDU.SG
Information Systems Technology and Design Pillar (ISTD)
Singapore University of Technology and Design, SUTD
Singapore 487372, 8 Somapah Road, Singapore

Ürün Dogan UDOGAN@MICROSOFT.COM
Microsoft Research
Cambridge CB1 2FB, UK

Marius Kloft KLOFT@HU-BERLIN.DE
Department of Computer Science
Humboldt University of Berlin
Berlin, Germany

Editors: Bob Durrant and Kee-Eung Kim

Abstract

We propose a localized approach to multiple kernel learning that can be formulated as a *convex* optimization problem over a given cluster structure. For which we obtain generalization error guarantees and derive an optimization algorithm based on the Fenchel dual representation. Experiments on real-world datasets from the application domains of computational biology and computer vision show that convex localized multiple kernel learning can achieve higher prediction accuracies than its global and non-convex local counterparts.
Keywords: Multiple kernel learning, Localized algorithms, Generalization analysis

1. Introduction

Kernel-based methods such as support vector machines have found diverse applications due to their distinct merits such as the descent computational complexity, high usability, and the solid mathematical foundation (e.g., Schölkopf and Smola, 2002). The performance of such algorithms, however, crucially depends on the involved kernel function as it intrinsically specifies the feature space where the learning process is implemented, and thus provides a similarity measure on the input space. Yet in the standard setting of these methods the choice of the involved kernel is typically left to the user.

A substantial step toward the complete automatization of kernel-based machine learning is achieved in Lanckriet et al. (2004), who introduce the *multiple kernel learning* (MKL)

framework (Gönen and Alpaydin, 2011). MKL offers a principal way of encoding complementary information with distinct base kernels and automatically learning an optimal combination of those (Sonnenburg et al., 2006a). MKL can be phrased as a single convex optimization problem, which facilitates the application of efficient numerical optimization strategies (Bach et al., 2004; Kloft et al., 2009; Sonnenburg et al., 2006a; Rakotomamonjy et al., 2008; Xu et al., 2010; Kloft et al., 2008a; Yang et al., 2011) and theoretical understanding of the generalization performance of the resulting models (Srebro and Ben-David, 2006; Cortes et al., 2010; Kloft et al., 2010; Kloft and Blanchard, 2011, 2012; Cortes et al., 2013; Ying and Campbell, 2009; Hussain and Shawe-Taylor, 2011; Lei and Ding, 2014). While early sparsity-inducing approaches failed to live up to its expectations in terms of improvement over uniform combinations of kernels (cf. Cortes, 2009, and references therein), it was shown that improved predictive accuracy can be achieved by employing appropriate regularization (Kloft et al., 2011, 2008b).

Currently, most of the existing algorithms fall into the *global* setting of MKL, in the sense that all input instances share the same kernel weights. However, this ignores the fact that instances may require sample-adaptive kernel weights.

For instance, consider the two images of a horses given to the right. Multiple kernels can be defined, capturing the shapes in the image and the color distribution over various channels. On the image to the left, the depicted horse and the image backgrounds exhibit distinctly different color distributions, while for the image to the right the contrary is the case. Hence, a color kernel is more significant to detect a horse in the image to the left than for the image the right. This example motivates studying localized approaches to MKL (Yang et al., 2009; Gönen and Alpaydin, 2008; Li et al., 2016; Lei et al., 2015; Mu and Zhou, 2011; Han and Liu, 2012).

Existing approaches to localized MKL (reviewed in Section 1.1) optimize *non-convex* objective functions. This puts their generalization ability into doubt. Indeed, besides the recent work by (Lei et al., 2015), the generalization performance of localized MKL algorithms (as measured through large-deviation bounds) is poorly understood, which potentially could make these algorithms prone to overfitting. Further potential disadvantages of non-convex localized MKL approaches include computationally difficulty in finding good local minima and the induced lack of reproducibility of results (due to varying local optima).

This paper presents a *convex* formulation of localized multiple kernel learning, which is formulated as a single convex optimization problem over a precomputed cluster structure, obtained through a potentially convex or non-convex clustering method. We derive an efficient optimization algorithm based on Fenchel duality. Using Rademacher complexity theory, we establish large-deviation inequalities for localized MKL, showing that the smoothness in the cluster membership assignments crucially controls the generalization error. Computational experiments on data from the domains of computational biology and computer vision show that the proposed convex approach can achieve higher prediction accuracies than its global and non-convex local counterparts (up to +5% accuracy for splice site detection).

1.1. Related Work

Gönen and Alpaydin (2008) initiate the work on localized MKL by introducing gating models

$$f(x) = \sum_{m=1}^{M} \eta_m(x; v)\langle w_m, \phi_m(x)\rangle + b, \quad \eta_m(x; v) \propto \exp(\langle v_m, x\rangle + v_{m0})$$

to achieve local assignments of kernel weights, resulting in a non-convex MKL problem. To not overly respect individual samples, Yang et al. (2009) give a group-sensitive formulation of localized MKL, where kernel weights vary at, instead of the example level, the group level. Mu and Zhou (2011) also introduce a non-uniform MKL allowing the kernel weights to vary at the cluster-level and tune the kernel weights under the graph embedding framework. Han and Liu (2012) built on Gönen and Alpaydin (2008) by complementing the spatial-similarity-based kernels with probability confidence kernels reflecting the likelihood of examples belonging to the same class. Li et al. (2016) propose a multiple kernel clustering method by maximizing *local* kernel alignments. Liu et al. (2014) present sample-adaptive approaches to localized MKL, where kernels can be switched on/off at the example level by introducing a latent binary vector for each individual sample, which and the kernel weights are then jointly optimized via margin maximization principle. Moeller et al. (2016) present a unified viewpoint of localized MKL by interpreting gating functions in terms of local reproducing kernel Hilbert spaces acting on the data. All the aforementioned approaches to localized MKL are formulated in terms of *non-convex* optimization problems, and deep theoretical foundations in the form of generalization error or excess risk bounds are unknown. Although Cortes et al. (2013) present a convex approach to MKL based on controlling the local Rademacher complexity, the meaning of *locality* is different in Cortes et al. (2013): it refers to the localization of the hypothesis class, which can result in sharper excess risk bounds (Kloft and Blanchard, 2011, 2012), and is not related to localized multiple kernel learning. Liu et al. (2015) extend the idea of sample-adaptive MKL to address the issue with missing kernel information on some examples. More recently, Lei et al. (2015) propose a MKL method by decoupling the locality structure learning with a hard clustering strategy from optimizing the parameters in the spirit of multi-task learning. They also develop the first generalization error bounds for localized MKL.

2. Convex Localized Multiple Kernel Learning

2.1. Problem setting and notation

Suppose that we are given n training samples $(x_1, y_1), \ldots, (x_n, y_n)$ that are partitioned into l disjoint clusters $S_1, \ldots, S_l$ in a probabilistic manner, meaning that, for each cluster S_j, we have a function $c_j : \mathcal{X} \to [0, 1]$ indicating the likelihood of x falling into cluster j, i.e., $\sum_{j\in\mathbb{N}_l} c_j(x) = 1$ for all $x \in \mathcal{X}$. Here, for any $d \in \mathbb{N}$, we introduce the notation $\mathbb{N}_d = \{1, \ldots, d\}$. Suppose that we are given M base kernels $k_1, \ldots, k_M$ with $k_m(x, \tilde{x}) = \langle \phi_m(x), \phi_m(\tilde{x})\rangle_{k_m}$, corresponding to linear models $f_j(x) = \langle w_j, \phi(x)\rangle + b = \sum_{m\in\mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x)\rangle + b$, where $w_j = (w_j^{(1)}, \ldots, w_j^{(M)})$ and $\phi = (\phi_1, \ldots, \phi_M)$. We consider the following proposed model, which is a weighted combination of these l local models:

$$f(x) = \sum_{j\in\mathbb{N}_l} c_j(x) f_j(x) = \sum_{j\in\mathbb{N}_l} c_j(x)\Big[\sum_{m\in\mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x)\rangle\Big] + b. \tag{1}$$

2.2. Proposed convex localized MKL method

The proposed convex localized MKL model can be formulated as follows.

Problem 1 (Convex Localized Multiple Kernel Learning (CLMKL)—Primal) *Let $C > 0$ and $p \geq 1$. Given a loss function $\ell(t, y) : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ convex w.r.t. the first argument and cluster likelihood functions $c_j : \mathcal{X} \to [0, 1]$, $j \in \mathbb{N}_l$, solve*

$$\begin{aligned} \inf_{w,t,\beta,b} & \sum_{j\in\mathbb{N}_l} \sum_{m\in\mathbb{N}_M} \frac{\|w_j^{(m)}\|_2^2}{2\beta_{jm}} + C \sum_{i\in\mathbb{N}_n} \ell(t_i, y_i) \\ s.t.\ \ & \beta_{jm} \geq 0, \sum_{m\in\mathbb{N}_M} \beta_{jm}^p \leq 1 \quad \forall j \in \mathbb{N}_l, m \in \mathbb{N}_M \\ & \sum_{j\in\mathbb{N}_l} c_j(x_i)\big[\sum_{m\in\mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x_i)\rangle\big] + b = t_i,\ \forall i \in \mathbb{N}_n. \end{aligned} \tag{P}$$

The core idea of the above problem is to use cluster likelihood functions for each example and separate ℓ_p-norm constraint on the kernel weights $\beta_j := (\beta_{j1}, \ldots, \beta_{jM})$ for each cluster j (Kloft et al., 2011) . Thus each instance can obtain separate kernel weights. The above problem is convex, since a quadratic over a linear function is convex (e.g., Boyd and Vandenberghe, 2004, p.g. 89). Note that Slater's condition can be directly checked, and thus strong duality holds.

2.3. Dualization

This section gives a dual representation of Problem 1. We consider two levels of duality: a partially dualized problem, with fixed kernel weights β_{jm}, and the entirely dualized problem with respect to all occurring primal variables. From the former we derive an efficient two-step optimization scheme (Section 3). The latter allows us to compute the duality gap and thus to obtain a sound stopping condition for the proposed algorithm. We focus on the entirely dualized problem here. The partial dualization is given in Appendix C of the long version of this paper (Lei et al., 2016).

Dual CLMKL Optimization Problem For $w_j = (w_j^{(1)}, \ldots, w_j^{(M)})$, we define the $\ell_{2,p}$-norm by $\|w_j\|_{2,p} := \|(\|w_j^{(1)}\|_{k_1}, \ldots, \|w_j^{(M)}\|_{k_M})\|_p = (\sum_{m\in\mathbb{N}_M} \|w_j^{(m)}\|_{k_m}^p)^{\frac{1}{p}}$. For a function h, we denote by $h^*(x) = \sup_\mu[x^\top \mu - h(\mu)]$ its Fenchel-Legendre conjugate. This results in the following dual.

Problem 2 (CLMKL—Dual) *The dual problem of (P) is given by*

$$\sup_{\sum_{i\in\mathbb{N}_n} \alpha_i = 0} \left\{ - C \sum_{i\in\mathbb{N}_n} \ell^*(-\frac{\alpha_i}{C}, y_i) - \frac{1}{2} \sum_{j\in\mathbb{N}_l} \Big\| \big(\sum_{i\in\mathbb{N}_n} \alpha_i c_j(x_i) \phi_m(x_i) \big)_{m=1}^M \Big\|_{2,\frac{2p}{p-1}}^2 \right\}. \tag{D}$$

Proof [**Dualization**] Using Lemma A.2 of (Lei et al., 2016) to express the optimal β_{jm} in terms of $w_j^{(m)}$, the problem (P) is equivalent to

$$\begin{aligned} &\inf_{w,t,b} \frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big(\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_2^{\frac{2p}{p+1}}\Big)^{\frac{p+1}{p}} + C\sum_{i\in\mathbb{N}_n}\ell(t_i,y_i) \\ &\text{s.t.}\ \sum_{j\in\mathbb{N}_l}\big[c_j(x_i)\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle\big] + b = t_i,\ \forall i\in\mathbb{N}_n. \end{aligned} \tag{2}$$

Introducing Lagrangian multipliers $\alpha_i, i\in\mathbb{N}_n$, the Lagrangian saddle problem of Eq. (2) is

$$\begin{aligned} &\sup_{\alpha}\inf_{w,t,b}\frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big(\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_2^{\frac{2p}{p+1}}\Big)^{\frac{p+1}{p}} + C\sum_{i\in\mathbb{N}_n}\ell(t_i,y_i) - \sum_{i\in\mathbb{N}_n}\alpha_i\Big(\sum_{j\in\mathbb{N}_l}c_j(x_i)\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle + b - t_i\Big) \\ &= \sup_{\alpha}\Big\{-C\sum_{i\in\mathbb{N}_n}\sup_{t_i}[-\ell(t_i,y_i) - \frac{1}{C}\alpha_i t_i] - \sup_b\sum_{i\in\mathbb{N}_n}\alpha_i b - \\ &\qquad \sup_w\Big[\sum_{j\in\mathbb{N}_l}\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i)\rangle - \frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big(\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_2^{\frac{2p}{p+1}}\Big)^{\frac{p+1}{p}}\Big]\Big\} \\ &\stackrel{\text{def}}{=} \sup_{\sum_{i\in\mathbb{N}_n}\alpha_i=0}\Big\{-C\sum_{i\in\mathbb{N}_n}\ell^*(-\frac{\alpha_i}{C},y_i) - \sum_{j\in\mathbb{N}_l}\Big[\frac{1}{2}\big\|\big(\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i)\big)_{m=1}^M\big\|_{2,\frac{2p}{p+1}}^2\Big]^*\Big\} \end{aligned} \tag{3}$$

The result (2) now follows by recalling that for a norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ is defined by $\|x\|_* = \sup_{\|\mu\|=1}\langle x,\mu\rangle$ and satisfies: $(\frac{1}{2}\|\cdot\|^2)^* = \frac{1}{2}\|\cdot\|_*^2$ (Boyd and Vandenberghe, 2004). Furthermore, it is straightforward to show that $\|\cdot\|_{2,\frac{2p}{p-1}}$ is the dual norm of $\|\cdot\|_{2,\frac{2p}{p+1}}$. ■

2.4. Representer Theorem

We can use the above derivation to obtain a lower bound on the optimal value of the primal optimization problem (P), from which we can compute the duality gap using the theorem below. The proof is given in Appendix A.2 in (Lei et al., 2016).

Theorem 3 (Representer Theorem) *For any dual variable $(\alpha_i)_{i=1}^n$ in (D), the optimal primal variable $\{w_j^{(m)}(\alpha)\}_{j,m=1}^{l,M}$ in the Lagrangian saddle problem (3) can be represented as*

$$w_j^{(m)}(\alpha) = \Big[\sum_{\tilde{m}\in\mathbb{N}_M}\|\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_{\tilde{m}}(x_i)\|_2^{\frac{2p}{p-1}}\Big]^{-\frac{1}{p}}\Big\|\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i)\Big\|_2^{\frac{2}{p-1}}\Big[\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i)\Big].$$

2.5. Support-Vector Classification

For the hinge loss, the Fenchel-Legendre conjugate becomes $\ell^*(t,y) = \frac{t}{y}$ (a function of t) if $-1 \le \frac{t}{y} \le 0$ and ∞ elsewise. Hence, for each i, the term $\ell^*(-\frac{\alpha_i}{C},y_i)$ translates to $-\frac{\alpha_i}{Cy_i}$, provided that $0 \le \frac{\alpha_i}{y_i} \le C$. With a variable substitution of the form $\alpha_i^{\text{new}} = \frac{\alpha_i}{y_i}$, the complete dual problem (D) reduces as follows.

Problem 4 (CLMKL—SVM Formulation) *For the hinge loss, the dual CLMKL problem* (D) *is given by:*

$$\sup_{\alpha:0\leq\alpha\leq C,\sum_{i\in\mathbb{N}_n}\alpha_i y_i=0} -\frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big\|\Big(\sum_{i\in\mathbb{N}_n}\alpha_i y_i c_j(x_i)\phi_m(x_i)\Big)_{m=1}^{M}\Big\|_{2,\frac{2p}{p-1}}^2+\sum_{i\in\mathbb{N}_n}\alpha_i, \tag{4}$$

A corresponding formulation for support-vector regression is given in Appendix B in (Lei et al., 2016).

3. Optimization Algorithms

As pioneered in Sonnenburg et al. (2006a), we consider here a two-layer optimization procedure to solve the problem (P) where the variables are divided into two groups: the group of kernel weights $\{\beta_{jm}\}_{j,m=1}^{l,M}$ and the group of weight vectors $\{w_j^{(m)}\}_{j,m=1}^{l,M}$. In each iteration, we alternatingly optimize one group of variables while fixing the other group of variables. These iterations are repeated until some optimality conditions are satisfied. To this aim, we need to find efficient strategies to solve the two subproblems.

It is not difficult to show (cf. Appendix C in (Lei et al., 2016)) that, given fixed kernel weights $\boldsymbol{\beta}=(\beta_{jm})$, the CLMKL dual problem is given by

$$\sup_{\boldsymbol{\alpha}:\sum_{i\in\mathbb{N}_n}\alpha_i=0} -\frac{1}{2}\sum_{j\in\mathbb{N}_l}\sum_{m\in\mathbb{N}_M}\beta_{jm}\Big\|\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i)\Big\|_2^2-C\sum_{i\in\mathbb{N}_n}\ell^*(-\frac{\alpha_i}{C},y_i), \tag{5}$$

which is a standard SVM problem using the kernel

$$\tilde{k}(x_i,x_{\tilde{i}}):=\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\beta_{jm}c_j(x_i)c_j(x_{\tilde{i}})k_m(x_i,x_{\tilde{i}}) \tag{6}$$

This allows us to employ very efficient existing SVM solvers (Chang and Lin, 2011). In the degenerate case with $c_j(x)\in\{0,1\}$, the kernel $\tilde{k}$ would be supported over those sample pairs belonging to the same cluster.

Next, we show that, the subproblem of optimizing the kernel weights for fixed $w_j^{(m)}$ and b has a closed-form solution.

Proposition 5 (Solution of the Subproblem w.r.t. the Kernel Weights) *Given fixed* $w_j^{(m)}$ *and* b*, the minimal* β_{jm} *in optimization problem* (P) *is attained for*

$$\beta_{jm}=\|w_j^{(m)}\|_2^{\frac{2}{p+1}}\Big(\sum_{k\in\mathbb{N}_M}\|w_j^{(k)}\|_2^{\frac{2p}{p+1}}\Big)^{-\frac{1}{p}}. \tag{7}$$

We present the detailed proof in Appendix A.3 in (Lei et al., 2016) due to lack of space. To apply Proposition 5 for updating β_{jm}, we need to compute the norm of $w_j^{(m)}$, and this can be accomplished by the following representation of $w_j^{(m)}$ given fixed β_{jm}: (cf. Appendix C in (Lei et al., 2016))

$$w_j^{(m)}=\beta_{jm}\sum_{i\in\mathbb{N}_n}\alpha_i c_j(x_i)\phi_m(x_i). \tag{8}$$

The prediction function is then derived by plugging the above representation into Eq. (1).

The resulting optimization algorithm for CLMKL is shown in Algorithm 1. The algorithm alternates between solving an SVM subproblem for fixed kernel weights (Line 4) and updating the kernel weights in a closed-form manner (Line 6). To improve the efficiency, we start with a crude precision and gradually improve the precision of solving the SVM subproblem. The proposed optimization approach can potentially be extended to an interleaved algorithm where the optimization of the MKL step is directly integrated into the SVM solver. Such a strategy can increase the computational efficiency by up to 1-2 orders of magnitude (cf. (Sonnenburg et al., 2006a) Figure 7 in Kloft et al. (2011)). The requirement to compute the kernel $\tilde{k}$ at each iteration can be further relaxed by updating only some randomly selected kernel elements.

Algorithm 1: Training algorithm for convex localized multiple kernel learning (CLMKL).

input: examples $\{(x_i, y_i)_{i=1}^n\} \subset \left(\mathcal{X} \times \{-1, 1\}\right)^n$ together with the likelihood functions $\{c_j(x)\}_{j=1}^l$, M base kernels $k_1, \ldots, k_M$.

initialize $\beta_{jm} = \sqrt[p]{1/M}, w_j^{(m)} = 0$ for all $j \in \mathbb{N}_l, m \in \mathbb{N}_M$

while *Optimality conditions are not satisfied* **do**

- calculate the kernel matrix $\tilde{k}$ by Eq. (6)
- compute α by solving canonical SVM with $\tilde{k}$
- compute $\|w_j^{(m)}\|_2^2$ for all j, m with $w_j^{(m)}$ given by Eq. (8)
- update β_{jm} for all j, m according to Eq. (7)

end

An alternative strategy would be to directly optimize (2) (without the need of a two-step wrapper approach). Such an approach has been presented in Sun et al. (2010) in the context of ℓ_p-norm MKL.

3.1. Convergence Analysis of the Algorithm

The theorem below, which is proved in Appendix A.4 in (Lei et al., 2016), shows convergence of Algorithm 1. The core idea is to view Algorithm 1 as an example of the classical block coordinate descent (BCD) method, convergence of which is well understood.

Theorem 6 (Convergence analysis of Algorithm 1) *Assume that*

(B1) the feature map $\phi_m(x)$ is of finite dimension, i.e, $\phi_m(x) \in \mathbb{R}^{e_m}, e_m < \infty, \forall m \in \mathbb{N}_M$
(B2) the loss function ℓ is convex, continuous w.r.t. the first argument and $\ell(0, y) < \infty, \forall y \in \mathcal{Y}$
(B3) any iterate β_{jm} traversed by Algorithm 1 has $\beta_{jm} > 0$
(B4) the SVM computation in line 4 of Algorithm 1 is solved exactly in each iteration.

Then, any limit point of the sequence traversed by Algorithm 1 minimizes the problem (P).

3.2. Runtime Complexity Analysis

At each iteration of the training stage, we need $O(n^2 M l)$ operations to calculate the kernel (6), $O(n^2 n_s)$ operations to solve a standard SVM problem, $O(M l n_s^2)$ operations to

calculate the norm according to the representation (8) and $O(Ml)$ operations to update the kernel weights. Thus, the computational cost at each iteration is $O(n^2Ml)$. The time complexity at the test stage is $O(n_t n_s Ml)$. Here, n_s and n_t are the number of support vectors and test points, respectively.

4. Generalization Error Bounds

In this section we present generalization error bounds for our approach. We give a purely data-dependent bound on the generalization error, which is obtained using Rademacher complexity theory (Bartlett and Mendelson, 2002). To start with, our basic strategy is to plug the optimal β_{jm} established in Eq. (7) into (P), so as to equivalently rewrite (P) as a block-norm regularized problem as follows:

$$\min_{w,b} \frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big[\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_2^{\frac{2p}{p+1}}\Big]^{\frac{p+1}{p}} + C\sum_{i\in\mathbb{N}_n}\ell\Big(\sum_{j\in\mathbb{N}_l}c_j(x_i)\big[\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle\big]+b,y_i\Big). \quad (9)$$

Solving (9) corresponds to empirical risk minimization in the following hypothesis space:

$$H_{p,D} := H_{p,D,M} = \Big\{f_w : x\to\sum_{j\in\mathbb{N}_l}c_j(x_i)\big[\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle\big] : \sum_{j\in\mathbb{N}_l}\|w_j\|^2_{2,\frac{2p}{p+1}}\le D\Big\}.$$

The following theorem establishes the Rademacher complexity bounds for the function class $H_{p,D}$, from which we derive generalization error bounds for CLMKL in Theorem 9. The proofs of the Theorems 8, 9 are given in Appendix A.5 in (Lei et al., 2016).

Definition 7 *For a fixed sample $S=(x_1,\ldots,x_n)$, the empirical Rademacher complexity of a hypothesis space H is defined as $\hat{R}_n(H) := \mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in H}\frac{1}{n}\sum_{i\in\mathbb{N}_n}\sigma_i f(x_i)$, where the expectation is taken w.r.t. $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_n)^\top$ with $\sigma_i, i\in\mathbb{N}_n$, being a sequence of independent uniform $\{\pm1\}$-valued random variables.*

Theorem 8 (CLMKL Rademacher complexity bounds) *The empirical Rademacher complexity of $H_{p,D}$ can be controlled by*

$$\hat{R}_n(H_{p,D}) \le \frac{\sqrt{D}}{n}\inf_{2\le t\le\frac{2p}{p-1}}\left(t\sum_{j\in\mathbb{N}_l}\Big\|\Big(\sum_{i\in\mathbb{N}_n}c_j^2(x_i)k_m(x_i,x_i)\Big)_{m=1}^M\Big\|_{\frac{t}{2}}\right)^{1/2}. \quad (10)$$

If, additionally, $k_m(x,x)\le B$ for any $x\in\mathcal{X}$ and any $m\in\mathbb{N}_M$, then we have

$$\hat{R}_n(H_{p,D}) \le \frac{\sqrt{DB}}{n}\inf_{2\le t\le\frac{2p}{p-1}}\Big(tM^{\frac{2}{t}}\sum_{j\in\mathbb{N}_l}\sum_{i\in\mathbb{N}_n}c_j^2(x_i)\Big)^{1/2}.$$

Theorem 9 (CLMKL Generalization Error Bounds) *Assume that $k_m(x,x)\le B, \forall m\in\mathbb{N}_M, x\in\mathcal{X}$. Suppose the loss function ℓ is L-Lipschitz and bounded by B_ℓ. Then, the following inequality holds with probability larger than $1-\delta$ over samples of size n for all classifiers $h\in H_{p,D}$:*

$$\mathcal{E}_\ell(h)\le\mathcal{E}_{\ell,\mathbf{z}}(h)+B_\ell\sqrt{\frac{\log(2/\delta)}{2n}}+2\frac{\sqrt{DB}}{n}\inf_{2\le t\le\frac{2p}{p-1}}\Big(tM^{\frac{2}{t}}\big[\sum_{j\in\mathbb{N}_l}\sum_{i\in\mathbb{N}_n}c_j^2(x_i)\big]\Big)^{1/2},$$

where $\mathcal{E}_\ell(h) := \mathbb{E}[\ell(h(x),y)]$ and $\mathcal{E}_{\ell,\mathbf{z}}(h) := \frac{1}{n}\sum_{i\in\mathbb{N}_n}\ell(h(x_i),y_i)$.

The above bound enjoys a mild dependence on the number of kernels. One can show (cf. Appendix A.5 in (Lei et al., 2016)) that the dependence is $O(\log M)$ for $p \leq (\log M - 1)^{-1} \log M$ and $O(M^{\frac{p-1}{2p}})$ otherwise. In particular, the dependence is logarithmically for $p = 1$ (sparsity-inducing CLMKL). These dependencies recover the best known results for global MKL algorithms in Cortes et al. (2010); Kloft and Blanchard (2011); Kloft et al. (2011).

The bounds of Theorem 8 exhibit a strong dependence on the likelihood functions, which inspires us to derive a new algorithmic strategy as follows. Consider the special case where $c_j(x)$ takes values in $\{0, 1\}$ (hard cluster membership assignment), and thus the term determining the bound has $\sum_{j\in\mathbb{N}_l} \sum_{m\in\mathbb{N}_M} c_j^2(x_i) = n$. On the other hand, if $c_j(x) \equiv \frac{1}{l}, j \in \mathbb{N}_l$ (uniform cluster membership assignment), we have the favorable term $\sum_{j\in\mathbb{N}} \sum_{i\in\mathbb{N}_n} c_j^2(x_i) = \frac{n}{l}$. This motivates us to introduce a parameter τ controlling the complexity of the bound by considering likelihood functions of the form

$$c_j(x) \propto \exp(-\tau \mathrm{dist}^2(x, S_j)), \tag{11}$$

where $\mathrm{dist}(x, S_j)$ is the distance between the example x and the cluster S_j. By letting $\tau = 0$ and $\tau = \infty$, we recover uniform and hard cluster assignments, respectively. Intermediate values of τ correspond to more balanced cluster assignments. As illustrated by Theorem 8, by tuning τ we optimally adjust the resulting models' complexities.

5. Empirical Analysis and Applications

5.1. Experimental Setup

We implement the proposed convex localized MKL (CLMKL) algorithm in MATLAB and solve the involved canonical SVM problem with LIBSVM (Chang and Lin, 2011). The clusters $\{S_1, \ldots, S_l\}$ are computed through kernel k-means (e.g., Dhillon et al., 2004), but in principle other clustering methods (including convex ones such as Hocking et al. (2011)) could be used. To further diminish k-means' potential fluctuations (which are due to random initialization of the cluster means), we repeat kernel k-means t times, and choose the one with minimal clustering error (the summation of the squared distance between the examples and the associated nearest cluster) as the final partition $\{S_1, \ldots, S_l\}$. To tune the parameter τ in (11) in a uniform manner, we introduce the notation

$$\mathrm{AE}(\tau) := \frac{1}{nl} \sum_{i\in\mathbb{N}_n} \sum_{j\in\mathbb{N}_l} \frac{\exp(-\tau \mathrm{dist}^2(x_i, S_j))}{\max_{\tilde{j}\in\mathbb{N}_l} \exp(-\tau \mathrm{dist}^2(x_i, S_{\tilde{j}}))}$$

to measure the average evenness (or average excess over hard partition) of the likelihood function. It can be checked that $\mathrm{AE}(\tau)$ is a strictly decreasing function of τ, taking value 1 at the point $\tau = 0$ and l^{-1} at the point $\tau = \infty$. Instead of tuning the parameter τ directly, we propose to tune the average excess/evenness over a subset in $[l^{-1}, 1]$. The associated parameter τ are then fixed by the standard binary search algorithm.

We compare the performance attained by the proposed CLMKL to regular localized MKL (LMKL) (Gönen and Alpaydin, 2008), localized MKL based on hard clustering (HLMKL) (Lei et al., 2015), the SVM using a uniform kernel combination (UNIF) (Cortes, 2009), and ℓ_p-norm MKL (Kloft et al., 2011), which includes classical MKL (Lanckriet et al.,

2004) as a special case. We optimize ℓ_p-norm MKL and CLMKL until the relative duality gap drops below 0.001. The calculation of the gradients in LMKL (Gönen and Alpaydin, 2008) requires $O(n^2M^2d)$ operations, which scales poorly, and the definition of the gating model requires the information of primitive features, which is not available for the biological applications studied below, all of which involve string kernels. In Appendix D of (Lei et al., 2016), we therefore give a fast and general formulation of LMKL, which requires only $O(n^2M)$ operations per iteration. Our implementation of which is available from the following webpage, together with our CLMKL implementation and scripts to reproduce the experiments: https://www.dropbox.com/sh/hkkfa0ghxzuig03/AADRdtSSdUSm8hfVbsdjcRqva?dl=0.

In the following we report detailed results for various real-world experiments. Further details are shown in Appendix E of (Lei et al., 2016).

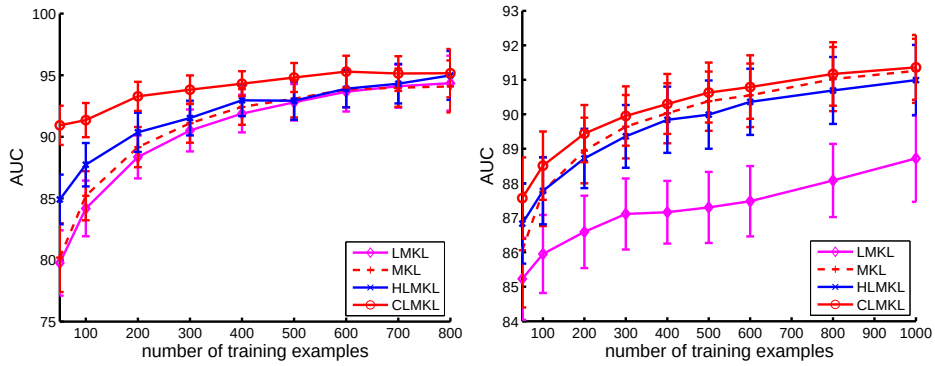
5.2. Splice Site Recognition

Our first experiment aims at detecting splice sites in the organism *Caenorhabditis elegans*, which is an important task in computational gene finding as splice sites are located on the DNA strang right at the boundary of exons (which code for proteins) and introns (which do not). We experiment on the `mkl-splice` data set, which we download from http://mldata.org/repository/data/viewslug/mkl-splice/. It includes 1000 splice site instances and 20 weighted-degree kernels with degrees ranging from 1 to 20 (Ben-Hur et al., 2008). The experimental setup for this experiment is as follows. We create random splits of this dataset into training set, validation set and test set, with size of training set traversing over the set $\{50, 100, 200, 300, \ldots, 800\}$. We apply kernel-kmeans with uniform kernel to generate a partition with $l = 3$ clusters for both CLMKL and HLMKL, and use this kernel to define the gating model in LMKL. To be consistent with previous studies, we use the area under the ROC curve (AUC) as an evaluation criterion. We tune the SVM regularization parameter from $10^{\{-1,-0.5,\ldots,2\}}$, and the average evenness over the interval $[0.4, 0.8]$ with eight linearly equally spaced points, based on the AUCs on the validation set. All the base kernel matrices are multiplicatively normalized before training. We repeat the experiment 50 times, and report mean AUCs on the test set as well as standard deviation. Figure 1 (a) shows the results as a function of the training set size n.

We observe that CLMKL achieves, for all n, a significant gain over all baselines. This improvement is especially strong for small n. For $n = 50$, CLMKL attains 90.9% accuracy, while the best baseline only achieves 85.4%, improving by 5.5%. Detailed results with standard deviation are reported in Table 1. A hypothetical explanation of the improvement from CLMKL is that splice sites are characterized by nucleotide sequences—so-called *motifs*—the length of which may differ from site to site (Sonnenburg et al., 2008). The 20 employed kernels count matching subsequences of length 1 to 20, respectively. For sites characterized by smaller motifs, low-degree WD-kernels are thus more effective than high-degree ones, and vice versa for sites containing longer motifs.

5.3. Transcription Start Site Detection

Our next experiment aims at detecting transcription start sites (TSS) of RNA Polymerase II binding genes in genomic DNA sequences. We experiment on the `TSS` data set, which we downloaded from http://mldata.org/repository/data/viewslug/tss/. This data set, which is included in the larger study of Sonnenburg et al. (2006b), comes with 5 kernels.



(a) `Splice` (b) `TSS`

Figure 1: Results of the gene finding experiments: splice site detection (left) and transcription start site detection (right). To clean the presentation, results for UNIF are not given here. The parameter p for CLMKL, HLMKL and MKL is set as 1 here.

	50	100	200	300	400	500	600	700	800
UNIF	79.5±2.8●	84.2±2.2●	88.0±1.7●	90.0±1.7●	91.6±1.5●	92.4±1.5●	93.3±1.7●	93.6±1.7●	93.8±2.3●
LMKL	79.8±2.7●	84.2±2.3●	88.4±1.7●	90.5±1.7●	91.9±1.5●	92.8±1.5●	93.7±1.6●	94.1±1.7●	94.3±2.2●
MKL,p=1	80.2±2.8●	85.2±2.0●	89.2±1.6●	91.1±1.6●	92.5±1.5●	93.1±1.5●	93.9±1.4●	94.0±1.6●	94.2±2.1●
MKL,p=2	79.6±2.8●	84.3±2.2●	88.3±1.7●	90.4±1.6●	91.8±1.5●	92.5±1.5●	93.4±1.6●	93.6±1.6●	93.8±2.2●
MKL,p=1.33	79.7±2.9●	84.6±2.1●	88.6±1.7●	90.6±1.6●	92.0±1.5●	92.7±1.5●	93.5±1.5●	93.7±1.6●	93.8±2.1●
HLMKL,p=1	84.9±2.0●	87.7±1.8●	90.4±1.6●	91.5±1.4●	93.0±1.3●	92.9±1.6●	93.9±1.5●	94.3±1.6●	95.0±2.0●
HLMKL,p=2	84.9±2.0●	87.0±1.7●	90.4±1.4●	91.1±1.6●	92.6±1.4●	93.5±1.6●	94.7±1.4●	94.6±1.4●	94.4±2.2●
HLMKL,p=1.33	85.4±1.9●	88.5±1.7●	90.1±1.6●	91.7±1.4●	92.7±1.2●	93.4±1.6●	94.6±1.5●	94.4±1.7●	94.4±2.1●
CLMKL,p=1	90.9±1.6	91.3±1.4●	93.3±1.2●	93.8±1.2●	94.3±1.0●	94.8±1.2	95.3±1.3●	95.1±1.4●	95.2±2.0●
CLMKL,p=2	90.5±1.6●	92.3±1.2●	93.0±1.2●	94.0±1.2	94.4±1.1●	94.7±1.2●	95.4±1.4●	95.3±1.5●	95.6±1.9●
CLMKL,p=1.33	90.9±1.5	90.1±1.3	92.7±1.2	94.1±1.2	94.8±1.1	94.9±1.1	95.6±1.2	95.4±1.5	95.4±1.9

Table 1: Performances achieved by LMKL, UNIF, regular ℓ_p MKL, HLMKL, and CLMKL on Splice Dataset. ● indicates that CLMKL with $p = 1.33$ is significantly better than the compared method (paired t-tests at 95% significance level).

The SVM based on the uniform combination of these 5 kernels was found to have the highest overall performance among 19 promoter prediction programs (Abeel et al., 2009). It therefore constitutes a strong baseline. To be consistent with previous studies (Abeel et al., 2009; Kloft, 2011; Sonnenburg et al., 2006b), we use the area under the ROC curve (AUC) as an evaluation criterion. We consider the same experimental setup as in the splice detection experiment. The gating function and the partition are computed with the TSS kernel, which carries most of the discriminative information (Sonnenburg et al., 2006b). All kernel matrices were normalized with respect to their trace, prior to the experiment.

Figure 1 (b) shows the AUCs on the test data sets as a function of the number of training examples.We observe that CLMKL attains a consistent improvement over other competing methods. Again, this improvement is most significant when n is small. Detailed results with standard deviation are reported in Table 2.

	50	100	200	300	400	500	600	800	1000
UNIF	83.9±2.4•	86.2±1.3•	87.6±1.0•	88.4±0.9•	88.7±0.9•	89.1±0.9•	89.2±1.0•	89.6±1.1•	89.8±1.1•
LMKL	85.2±1.2•	85.9±1.1•	86.6±1.1•	87.1±1.0•	87.2±0.9•	87.3±1.0•	87.5±1.0•	88.1±1.1•	88.7±1.3•
MKL,p=1	86.0±1.7•	87.7±1.0•	88.9±0.9•	89.6±0.9•	90.0±0.9•	90.3±0.9•	90.5±0.9	91.0±0.9	91.2±0.9
MKL,p=2	85.1±2.0•	86.9±1.1•	88.1±0.9•	88.8±0.9•	89.2±0.9•	89.6±0.9•	89.8±1.0•	90.3±1.0•	90.7±0.9•
MKL,p=1.33	85.7±1.8•	87.5±1.0•	88.7±0.9•	89.4±0.9•	89.8±0.9•	90.2±0.9•	90.4±0.9•	90.9±0.9•	91.2±0.9•
HLMKL,p=1	86.8±1.2•	87.8±1.0•	88.7±0.9•	89.4±0.9•	89.8±1.0•	90.0±1.0•	90.4±1.0•	90.7±1.0•	91.0±1.0•
HLMKL,p=2	86.3±1.4•	87.5±1.0•	88.5±0.9•	89.3±0.9•	89.4±0.9•	89.7±0.9•	89.8±1.0•	90.3±1.1•	90.5±1.0•
HLMKL,p=1.33	86.5±1.4•	87.7±1.1•	88.7±0.9•	89.3±0.9•	89.8±1.0•	90.1±0.9•	90.2±1.0•	90.7±1.0•	91.0±0.9•
CLMKL,p=1	87.6±1.2	88.5±1.0	89.4±0.8	90.0±0.9	90.3±0.9•	90.6±0.9	90.8±0.9•	91.2±0.9•	91.4±0.9
CLMKL,p=2	87.3±1.3•	88.3±1.0•	89.1±0.8•	89.6±0.8•	89.9±0.9•	90.2±0.9•	90.3±0.9•	90.7±1.0•	90.9±0.9•
CLMKL,p=1.33	87.6±1.2	88.6±0.9	89.4±0.8	89.9±0.9	90.2±0.9	90.5±0.9	90.6±1.0	91.1±1.0	91.3±0.9

Table 2: Performances achieved by LMKL, UNIF, regular ℓ_p MKL, HLMKL and CLMKL on TSS Dataset. • indicates that CLMKL with $p = 1.33$ is significantly better than the compared method (paired t-tests at 95% significance level).

5.4. Protein Fold Prediction

Protein fold prediction is a key step towards understanding the function of proteins, as the folding class of a protein is closely linked with its function; thus it is crucial for drug design. We experiment on the protein folding class prediction dataset by Ding and Dubchak (2001), which was also used in Campbell and Ying (2011); Kloft (2011); Kloft and Blanchard (2011). This dataset consists of 27 fold classes with 311 proteins used for training and 383 proteins for testing. We use exactly the same 12 kernels as in Campbell and Ying (2011); Kloft (2011); Kloft and Blanchard (2011) reflecting different features, such as van der Waals volume, polarity and hydrophobicity. We precisely replicate the experimental setup of previous experiments by Campbell and Ying (2011); Kloft (2011); Kloft and Blanchard (2011), which is detailed in Appendix E.1 of (Lei et al., 2016). We report the mean prediction accuracies, as well as standard deviations in Table 3.

The results show that CLMKL surpasses regular ℓ_p-norm MKL for all values of p, and achieves accuracies up to 0.6% higher than the one reported in Kloft (2011), which is higher than the initially reported accuracies in Campbell and Ying (2011). LMKL works poorly in this dataset, possibly because LMKL based on precomputed custom kernels requires to optimize nM additional variables, which may overfit.

5.5. Visual Image Categorization—UIUC Sports

We experiment on the UIUC Sports event dataset (Li and Fei-Fei, 2007) consisting of 1574 images, belonging to 8 image classes of sports activities. We compute 9 χ^2-kernels based

	UNIF	LMKL	MKL			HLMKL			CLMKL		
			$p=1$	$p=1.2$	$p=2$	$p=1$	$p=1.2$	$p=2$	$p=1$	$p=1.2$	$p=2$
ACC	68.4•	64.3•	68.7•	74.2•	70.8•	72.7 ± 1.3•	74.6 ± 0.6	72.4 ± 0.8•	71.3 ± 0.5•	**75.0 ± 0.7**	71.7 ± 0.5•

Table 3: Results of the protein fold prediction experiment. • indicates that CLMKL with $p = 1.2$ is significantly better than the compared method (paired t-tests at 95% significance level).

on SIFT features and global color histograms, which is described in detail in Appendix E.2 of (Lei et al., 2016), where we also give background on the experimental setup.

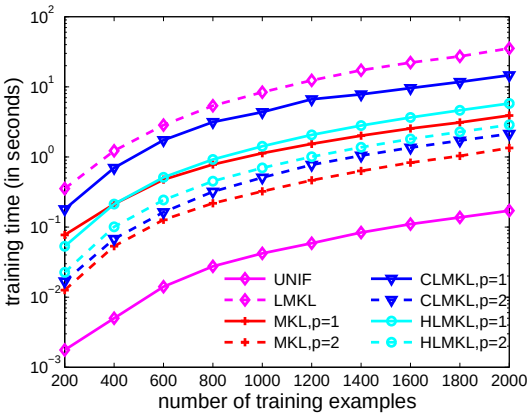
From the results shown in Table 4, we observe that CLMKL achieves a performance improvement by 0.26% over the ℓ_p-norm MKL baseline while localized MKL as in Gönen and Alpaydin (2008) underperforms the MKL baseline.

	MKL	LMKL	CLMKL		MKL	LMKL	CLMKL
ACC	90.00	87.29	90.26	Δ	0+ 11= 0-	0+ 1= 10-	4+ 6= 1-

Table 4: Results of the visual image recognition experiment on the UIUC sports dataset. Δ indicates on how many outer cross validation test splits a method is worse ($n-$), equal ($n=$) or better ($n+$) than MKL.

5.6. Execution Time Experiments

To demonstrate the efficiency of the proposed implementation, we compare the training time for UNIF, LMKL, ℓ_p-norm MKL, HLMKL and CLMKL on the TSS dataset. We fix the regularization parameter $C = 1$. We fix $l = 3$ and AE $= 0.5$ for CLMKL, and fix $l = 3$ for HLMKL. On the image to the right, we plot the training time versus the training set size. We repeat the experiment 20 times and report the average training time here. We optimize CLMKL, HLMKL and MKL until the relative gap is under 10^{-3}. The figure implies that CLMKL converges faster than LMKL. Furthermore, training an ℓ_2-norm MKL requires significantly less time than training an ℓ_1-norm MKL, which is consistent with the fact that the dual problem of ℓ_2-norm MKL is much smoother than the ℓ_1-norm counterpart.



6. Conclusions

Localized approaches to multiple kernel learning allow for flexible distribution of kernel weights over the input space, which can be a great advantage when samples require varying kernel importance. As we show in this paper, this can be the case in image recognition and several computational biology applications. However, almost prevalent approaches to localized MKL require solving difficult non-convex optimization problems, which makes them potentially prone to overfitting as theoretical guarantees such as generalization error bounds are yet unknown.

In this paper, we propose a theoretically grounded approach to localized MKL, consisting of two subsequent steps: 1. clustering the training instances and 2. computation of the kernel weights for each cluster through a single convex optimization problem. For which we derive an efficient optimization algorithm based on Fenchel duality. Using Rademacher complexity theory, we establish large-deviation inequalities for localized MKL, showing that

the smoothness in the cluster membership assignments crucially controls the generalization error. The proposed method is well suited for deployment in the domains of computer vision and computational biology. For splice site detection, CLMKL achieves up to 5% higher accuracy than its global and non-convex localized counterparts.

Future work could analyze extension of the methodology to semi-supervised learning (Görnitz et al., 2009, 2013) or using different clustering objectives (Vogt et al., 2015; Hocking et al., 2011) and how to principally include the construction of the data partition into our framework by constructing partitions that can capture the local variation of prediction importance of different features.

Acknowledgments

YL acknowledges support from the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006]. AB acknowledges support from Singapore University of Technology and Design Startup Grant SRIS15105. MK acknowledges support from the German Research Foundation (DFG) award KL 2698/2-1 and from the Federal Ministry of Science and Education (BMBF) awards 031L0023A and 031B0187B.

References

Thomas Abeel, Yves Van de Peer, and Yvan Saeys. Toward a gold standard for promoter prediction evaluation. *Bioinformatics*, 25(12):i313–i320, 2009.

Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, page 6, 2004.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4, 2008.

Stephen Poythress Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge Univ. Press, New York, 2004.

Colin Campbell and Yiming Ying. Learning with support vector machines. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(1):1–95, 2011.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Corinna Cortes. Invited talk: Can learning kernels help performance? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1:1–1:1, New York, NY, USA, 2009. ACM.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'10, 2010.

Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems*, pages 2760–2768, 2013.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM, 2004.

Chris HQ Ding and Inna Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

Mehmet Gönen and Ethem Alpaydin. Localized multiple kernel learning. In *Proceedings of the 25th international conference on Machine learning*, pages 352–359. ACM, 2008.

Mehmet Gönen and Ethem Alpaydin. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011. ISSN 1532-4435.

Nico Görnitz, Marius Kloft, and Ulf Brefeld. Active and semi-supervised data domain description. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–422. Springer, 2009.

Nico Görnitz, Marius Micha Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 2013.

Yina Han and Guizhong Liu. Probability-confidence-kernel-based localized multiple kernel learning with norm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(3):827–837, 2012.

Toby Dylan Hocking, Armand Joulin, Francis Bach, Jean-Philippe Vert, et al. Clusterpath: an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, 2011.

Zakria Hussain and John Shawe-Taylor. Improved loss bounds for multiple kernel learning. In *AISTATS*, pages 370–377, 2011.

Marius Kloft. *ℓ_p-norm multiple kernel learning*. PhD thesis, Berlin Institute of Technology, 2011.

Marius Kloft and Gilles Blanchard. The local Rademacher complexity of ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 24*, pages 2438–2446. MIT Press, 2011.

Marius Kloft and Gilles Blanchard. On the convergence rate of lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 13(1):2465–2502, 2012.

Marius Kloft, Ulf Brefeld, Patrick Düessel, Christian Gehl, and Pavel Laskov. Automatic feature selection for anomaly detection. In *Proceedings of the 1st ACM workshop on Workshop on AISec*, pages 71–76. ACM, 2008a.

Marius Kloft, Ulf Brefeld, Pavel Laskov, and Sören Sonnenburg. Non-sparse multiple kernel learning. In *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, volume 4, 2008b.

Marius Kloft, Ulf Brefeld, Pavel Laskov, Klaus-Robert Müller, Alexander Zien, and Sören Sonnenburg. Efficient and accurate lp-norm multiple kernel learning. In *Advances in neural information processing systems*, pages 997–1005, 2009.

Marius Kloft, Ulrich Rückert, and Peter Bartlett. A unifying view of multiple kernel learning. *Machine Learning and Knowledge Discovery in Databases*, pages 66–81, 2010.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

Yunwen Lei and Lixin Ding. Refined Rademacher chaos complexity bounds with applications to the multikernel learning problem. *Neural. Comput.*, 26(4):739–760, 2014.

Yunwen Lei, Alexander Binder, Ürün Dogan, and Marius Kloft. Theory and algorithms for the localized setting of learning kernels. In *Proceedings of The 1st International Workshop on Feature Extraction: Modern Questions and Challenges, NIPS*, pages 173–195, 2015.

Yunwen Lei, Alexander Binder, Ürün Dogan, and Marius Kloft. Localized multiple kernel learning—a convex approach. *arXiv preprint arXiv:1506.04364v2*, 2016.

Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE,

2007.

Miaomiao Li, Xinwang Liu, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel clustering with local kernel alignment maximization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, 2016.

Xinwang Liu, Lei Wang, Jian Zhang, and Jianping Yin. Sample-adaptive multiple kernel learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 1975–1981, 2014.

Xinwang Liu, Lei Wang, Jianping Yin, Yong Dou, and Jian Zhang. Absent multiple kernel learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2807–2813, 2015.

John Moeller, Sarathkrishna Swaminathan, and Suresh Venkatasubramanian. A unified view of localized kernel learning. *arXiv preprint arXiv:1603.01374*, 2016.

Yadong Mu and Bingfeng Zhou. Non-uniform multiple kernel learning with cluster-based gating functions. *Neurocomputing*, 74(7):1095–1101, 2011.

Alain Rakotomamonjy, Francis Bach, Stéphane Canu, Yves Grandvalet, et al. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006a.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. Arts: accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–e480, 2006b.

Sören Sonnenburg, Alexander Zien, Petra Philips, and Gunnar Rätsch. Poims: positional oligomer importance matricesunderstanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–i14, 2008.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *COLT*, pages 169–183. Springer-Verlag, Berlin, 2006.

Zhaonan Sun, Nawanol Ampornpunt, Manik Varma, and Svn Vishwanathan. Multiple kernel learning and the smo algorithm. In *Advances in neural information processing systems*, pages 2361–2369, 2010.

Julia E Vogt, Marius Kloft, Stefan Stark, Sudhir S Raman, Sandhya Prabhakaran, Volker Roth, and Gunnar Rätsch. Probabilistic clustering of time-evolving distance data. *Machine Learning*, 100 (2-3):635–654, 2015.

Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1175–1182, 2010.

Haiqin Yang, Zenglin Xu, Jieping Ye, Irwin King, and Michael R Lyu. Efficient sparse generalized multiple kernel learning. *IEEE Transactions on Neural Networks*, 22(3):433–446, 2011.

Jingjing Yang, Yuanning Li, Yonghong Tian, Lingyu Duan, and Wen Gao. Group-sensitive multiple kernel learning for object categorization. In *2009 IEEE 12th International Conference on Computer Vision*, pages 436–443. IEEE, 2009.

Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *COLT*, 2009.