JMLR: Workshop and Conference Proceedings 44 (2015) 173-195 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

Theory and Algorithms for the Localized Setting of Learning Kernels

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

Alexander Binder ALEXANDER_BINDER@SUTD.EDU.SG
Machine Learning Group, TU Berlin
ISTD Pillar, Singapore University of Technology and Design

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

Editor: Dmitry Storcheus

Abstract

We analyze the localized setting of learning kernels also known as localized multiple kernel learning. This problem has been addressed in the past using rather heuristic approaches based on approximately optimizing non-convex problem formulations, of which up to now no theoretical learning bounds are known. In this paper, we show generalization error bounds for learning localized kernel classes where the localities are coupled using graph-based regularization. We propose a novel learning localized kernels algorithm based on this hypothesis class that is formulated as a convex optimization problem using a pre-obtained cluster structure of the data. We derive dual representations using Fenchel conjugation theory, based on which we give a simple yet efficient wrapper-based optimization algorithm. We apply the method to problems involving multiple heterogeneous data sources, taken from domains of computational biology and computer vision. The results show that the proposed convex approach to learning localized kernels can achieve higher prediction accuracies than its global and non-convex local counterparts.

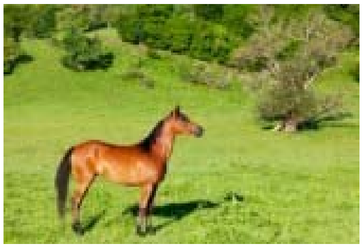
1. Introduction

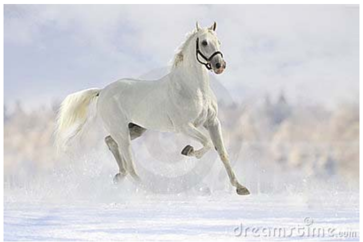
Kernel-based learning algorithms (e.g., Schölkopf and Smola, 2002) including support vector machines (Cortes and Vapnik, 1995) have found diverse applications due to their distinct merits such as decent computational complexity, high prediction accuracy (Delgado et al., 2014), and solid mathematical foundation (e.g., Mohri et al., 2012). Since the learning and data representation processes are decoupled in a modular fashion, one can obtain non-linear kernel machines from simpler linear ones in a canonical way. The performance of such algorithms, however, is fundamentally limited through the choice of the involved kernel function as it intrinsically specifies the feature space where the learning process is implemented. This choice is typically left to the user. A substantial step to-

©2015 Yunwen Lei, Alexander Binder, Ürün Dogan and Marius Kloft.

ward the complete automatization of kernel-based machine learning is achieved in Lanckriet et al. (2004), who introduce the *multiple kernel learning* (MKL) or *learning kernels* framework (Gönen and Alpaydin, 2011). Being formulated in terms of a single convex optimization criterion, MKL offers a theoretically sound way (Wu et al., 2007; Ying and Campbell, 2009; Cortes et al., 2010; Kloft and Blanchard, 2011, 2012; Cortes et al., 2013; Lei and Ding, 2014) of encoding complementary information with distinct base kernels and automatically learning an optimal combination of those (Ben-Hur et al., 2008; Gehler and Nowozin, 2008) using efficient numerical algorithms (Bach et al., 2004; Sonnenburg et al., 2006; Rakotomamonjy et al., 2008). This is particularly significant in the application domains of bioinformatics and computer vision (Ben-Hur et al., 2008; Gehler and Nowozin, 2008; Kloft, 2011), where data can be obtained from multiple heterogeneous sources, describing different properties of one and the same object (e.g., genome or image). 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).

Currently, most of the existing algorithms fall into the *global* setting of MKL, in the sense that the kernel combination is not varied over the input space. This ignores the fact that different regions of the input space might require individual kernel weights. For instance in the figures to the right, the images exhibit very distinct color distributions. While a kernel based on global color histograms may be effective to detect the *horse* object on the image to the left, it may fail in the image to the right, as the image fore- and backgrounds exhibit very similar color distributions. This motivates us studying *localized* approaches to learning kernels (Gönen and Alpaydin, 2008). The existing algorithms (reviewed in the subsequent section), however, optimize *non-convex* objective functions using ad-hoc optimization heuristics, which confuses the issue of reproducibility. Whether or not these algorithms are protected against *overfitting* is still an open research question as no theoretical guarantees—neither generalization error nor excess risk bounds—are known.





In this paper, we show generalization error bounds for a localized setting of learning kernels, where we assume a pre-specified cluster structure of the data. We show that performing empirical risk minimization over this class is given by a *convex* optimization problem. For which we derive partial and complete dual representations using Fenchel conjugation theory and derive an efficient convex wrapper-based optimization algorithm. We apply the method to problems involving multiple heterogeneous data sources, taken from domains of computational biology and computer vision. The results show that the proposed convex approach to learning localized kernels can achieve higher prediction accuracies than its global and non-convex local counterparts.

The remainder of this paper is structured as follows. In Section 2 we review related work; in Section 3 our convex and localized formulation of learning kernels is introduced, a partial dual representation of which is derived in Section 4, where we also present an efficient optimization algorithm. We report on theoretical results including generalization error bounds in Section 5. Empirical results for the application domains of visual image recognition and protein fold class prediction are presented in Section 6; Section 7 concludes.

2. Related work

Gönen and Alpaydin (2008) initiated the work on localized MKL by using a discriminant function $f(x) = \sum_{k=1}^{M} \eta_k(x|V)\langle w_k, \phi_k(x)\rangle + b$, where M is the number of kernels, $\eta_k(x|V)$ is a parametric gating model assigning a weight to $\phi_k(x)$ as a function of x, and V encodes the parameters of the gating model. The gating function is used to divide the input space into different regions, each of which is assigned to kernel weights. The joint optimization of the gating model and the kernel-based prediction function is carried out by alternating optimization. This problem is non-convex due to the non-linearity introduced by the gating function. Yang et al. (2009) develop a group-sensitive variant of MKL tailored to object categorization. Their approach is non-convex but, in contrast to Gönen and Alpaydin (2008), examples within a group share the same kernel weights while examples from different groups employ distinct sets of kernel weights. Han and Liu (2012) modify the approach of Gönen and Alpaydin (2008) by complementing the spatial-similarity-based kernels with probability confidence kernels that reflect the degree of confidence to which the involved examples belong to the same class. Song et al. (2011) present a localized MKL algorithm for realistic human action recognition in videos. However, the involved local models are constructed in an independent fashion. Therefore, they ignore the coupling among different localities, and may produce a suboptimal classifier already when these localities are moderately correlated. Recently, a localized MKL formulation has been studied as a computational means to study non-linear SVMs (Jose et al., 2013).

All these approaches are based on non-convex optimization criteria and lack in learning theory. To our knowledge, the only theoretically sound approach in the context of the localized setting of learning kernels is by Cortes, Kloft, and Mohri (2013). They present an MKL approach based on controlling the local Rademacher complexity of the resulting kernel combination. Note that the meaning of *locality* is different here, however: while in the present work we perform assignments of kernel weights locally with respect to the input space, Cortes, Kloft, and Mohri (2013) localize the hypothesis class, which leads to sharper generalization bounds (Kloft and Blanchard, 2011, 2012).

3. Learning methodology

In this paper, we study a convex formulation of localized MKL (CLMKL). For simplicity, we present our approach for binary classification, although the approach is general and can be extended to regression, multi-class classification, and structured output prediction.

3.1 Localized Problem Setting of Learning Kernels

Suppose we are given M base kernels $k_1, \ldots, k_M$ with ϕ_m being the corresponding kernel feature map corresponding to m-th kernel, i.e., $k_m(x, \tilde{x}) = \langle \phi_m(x), \phi_m(\tilde{x})\rangle_{k_m}$. Let $\mathcal{H}_m$ be the reproducing kernel Hilbert space corresponding to kernel k_m, inner product $\langle \cdot, \cdot \rangle_{k_m}$ and induced norm $\| \cdot \|_{k_m}$. For clarity, we frequently use the notation $\langle \cdot, \cdot \rangle := \langle \cdot, \cdot \rangle_{k_m}$ and $\| \cdot \|_2 := \| \cdot \|_{k_m}$. For any $d \in \mathbb{N}^+$, introduce the notation $\mathbb{N}_d = \{1, \ldots, d\}$. Suppose that the training examples $(x_1, y_1), \ldots, (x_n, y_n)$ are partitioned into l disjoint clusters $S_1, \ldots, S_l$. For each cluster S_j, we learn a distinct linear combined kernel $\tilde{k}_j = \sum_{m \in \mathbb{N}_M} \beta_{jm} k_m$ and a distinct weight vector $w_j = (w_j^{(1)}, \ldots, w_j^{(M)})$. This results, for each cluster S_j, in a linear model $f_j(x) = \langle w_j, \phi(x)\rangle + b = \sum_{m \in \mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x)\rangle + b$, where $\phi = (\phi_1, \ldots, \phi_M)$ is the concatenated feature map.

3.2 Notation

For a Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$ and l elements $w_1,\ldots,w_l\in\mathcal{H}$, we define the Σ semi-norm for $(w_1,\ldots,w_l)$ by

$$\|(w_1,\ldots,w_l)\|_\Sigma := \Big(\sum_{j,\tilde{j}\in\mathbb{N}_l}\Sigma_{j\tilde{j}}\langle w_j,w_{\tilde{j}}\rangle\Big)^{1/2}, \tag{1}$$

where Σ is a positive semi-definite $l\times l$ matrix. For any $\beta=(\beta_{jm})_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}$ and any $m\in\mathbb{N}_M$, we write $Q_m^\beta := Q_m^{\beta,\frac{1}{\mu}\Sigma}=(q^{(\beta)}_{mj\tilde{j}})_{j,\tilde{j}\in\mathbb{N}_l}=[\mathrm{diag}(\beta_{1m}^{-1},\ldots,\beta_{lm}^{-1})+\mu\Sigma^{-1}]^{-1}$, where $\mathrm{diag}(a_1,\ldots,a_l)$ is the $l\times l$ diagonal matrix with $a_1,\ldots,a_l$ on the main diagonal. For any $x\in\mathcal{X}$, we use $\tau(x)$ to denote the index of the cluster to which the point x belongs, i.e., $\tau(x)=j \iff x\in S_j$. For brevity, we write $\tau(i):=\tau(x_i)$ for all i and $a_+=\max(a,0)$ for all $a\in\mathbb{R}$. Introduce the notation $w^{(m)}=\big(w_1^{(m)},\ldots,w_l^{(m)}\big)$. For any $p\geq 1$, we denote by p^* its conjugated exponent, satisfying $\frac{1}{p}+\frac{1}{p^*}=1$. For $w_j=(w_j^{(1)},\ldots,w_j^{(M)})$, we define the $\ell_{2,p}$-norm by $\|w_j\|_{2,p}:=\Big(\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_{k_m}^p\Big)^{1/p}$.

3.3 Convex localized multiple kernel learning (CLMKL)

The proposed convex formulation for localized MKL is given as follows (for simplicity presented in terms of the hinge loss function; for a general presentation, see Appendix B.2):

Problem 1 (CONVEX LOCALIZED MULTIPLE KERNEL LEARNING (CLMKL))

$$\begin{aligned}\min_{w,\xi,\beta,b}\ & \sum_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}\frac{\|w_j^{(m)}\|_2^2}{2\beta_{jm}}+\frac{\mu}{2}\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|^2_{\Sigma^{-1}}+C\sum_{i\in\mathbb{N}_n}\xi_i\\ \textit{s.t.}\ & y_i\Big(\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle+b\Big)\geq 1-\xi_i,\ \xi_i\geq 0,\forall i\in S_j,j\in\mathbb{N}_l\\ & \sum_{m\in\mathbb{N}_M}\beta_{jm}^p\leq 1,\ \forall j\in\mathbb{N}_l,\beta_{jm}\geq 0,\ \forall j\in\mathbb{N}_l,m\in\mathbb{N}_M,\end{aligned} \tag{2}$$

where ξ_i are slack variables, C and μ are regularization parameters and Σ^{-1} is a positive semi-definite matrix (note that we do not need to compute the inversion of Σ^{-1} in the implementations).

Note that, we impose, for each cluster $S_j, j\in\mathbb{N}_l$, a separate ℓ_p-norm constraint (Kloft et al., 2011) on the combination coefficients $\beta_j=(\beta_{j1},\ldots,\beta_{jM})$. However, unlike training a local model independently at each locality, these l local models are optimized jointly in our formulation, exploiting that examples in localities close by may convey complementary information to the learning task. The regularizer defined in (1) encodes the relationship among different clusters and imposes a soft constraint on how these local models shall be correlated. Note that, if Σ^{-1} is the graph Laplacian of an adjacency matrix W (i.e., $\Sigma^{-1}=D-W$ with $D_{j\tilde{j}}=\delta_{j\tilde{j}}\sum_{k\in\mathbb{N}_l}W_{jk}$), the regularizer (1) coincides with the graph regularizer employed also in Evgeniou et al. (2005): $\|w^{(m)}\|^2_{\Sigma^{-1}}=\sum_{j,\tilde{j}\in\mathbb{N}_l}W_{j\tilde{j}}\|w_j^{(m)}-w_{\tilde{j}}^{(m)}\|_2^2$. Recall that a quadratic over a linear function is convex (e.g., Boyd and Vandenberghe, 2004, p.g. 89), so all occurring summands in formulation (2) are convex, so this is a convex optimization problem. Note that Slater's condition can be directly checked, and thus strong duality holds. *To our best knowledge, problem* (2) *is the first convex formulation in the localized setting of learning kernels.*

4. Optimization Algorithms

As pioneered in Sonnenburg et al. (2006), we consider here a two-layer optimization procedure to solve the problem (2) where the variables are divided into two groups: the group of kernel weights $\{\beta_{jm}\}_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}$ and the group of weight vectors $\{w_j^{(m)}\}_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}$. 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. The following proposition indicates that the subproblem of optimizing the objective of (2) with respect to $\{w_j^{(m)}\}_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}$ for fixed kernel weights can be cast as a standard SVM problem with a delicately defined kernel.

Proposition 2 (CLMKL (PARTIAL) DUAL PROBLEM) *Introduce the kernel*

$$\tilde{k}(x_i,x_{\tilde{i}}):=\sum_{m\in\mathbb{N}_M} q^{(\beta)}_{m\tau(i)\tau(\tilde{i})}k_m(x_i,x_{\tilde{i}}). \tag{3}$$

The partial Lagrangian dual of (2) *with fixed kernel weights* β_{jm} *is given by*

$$\begin{aligned}\max_{\alpha_i}\quad & \sum_{i\in\mathbb{N}_n}\alpha_i-\frac{1}{2}\sum_{i,\tilde{i}\in\mathbb{N}_n} y_iy_{\tilde{i}}\alpha_i\alpha_{\tilde{i}}\tilde{k}(x_i,x_{\tilde{i}})\\ \textit{s.t.}\quad & \sum_{i\in\mathbb{N}_n}\alpha_iy_i=0,\ 0\le\alpha_i\le C,\ \forall i\in\mathbb{N}_n.\end{aligned} \tag{4}$$

Further, the optimal weight vector can be represented by

$$w_j^{(m)}=\sum_{\tilde{j}\in\mathbb{N}_l} q^{(\beta)}_{mj\tilde{j}}\sum_{i\in S_{\tilde{j}}} y_i\alpha_i\phi_m(x_i),\qquad \forall j\in\mathbb{N}_l, m\in\mathbb{N}_M. \tag{5}$$

Next, we show that, the subproblem of optimizing the kernel weights for fixed $w_j^{(m)}$ and b has a closed-form solution. We defer the detailed proof of Propositions 2, 3 to the appendix due to the lack of the space.

Proposition 3 (SOLVING THE SUBPROBLEM WITH RESPECT TO THE KERNEL WEIGHTS) *Given fixed* $w_j^{(m)}$ *and b, the minimal* β_{jm} *in optimization problem* (2) *is attained for*

$$\beta_{jm}=\|w_j^{(m)}\|_2^{\frac{2}{p+1}}\Big(\sum_{k\in\mathbb{N}_M}\|w_j^{(k)}\|_2^{\frac{2p}{p+1}}\Big)^{-\frac{1}{p}}. \tag{6}$$

To apply Proposition 3 for updating β_{jm}, we need to compute the norm of $w_j^{(m)}$, which can be accomplished by recalling the representation given in Eq. (5):

$$\|w_j^{(m)}\|_2^2=\Big\|\sum_{i\in\mathbb{N}_n} y_i\alpha_i q^{(\beta)}_{mj\tau(i)}\phi_m(x_i)\Big\|_2^2=\sum_{i\in\mathbb{N}_n}\sum_{\tilde{i}\in\mathbb{N}_n} y_iy_{\tilde{i}}\alpha_i\alpha_{\tilde{i}}q^{(\beta)}_{mj\tau(i)}q^{(\beta)}_{mj\tau(\tilde{i})}k_m(x_i,x_{\tilde{i}}). \tag{7}$$

Furthermore, note that the prediction function becomes

$$f(x)=\sum_{m\in\mathbb{N}_M}\langle w^{(m)}_{\tau(x)},\phi_m(x)\rangle+b=\sum_{i\in\mathbb{N}_n} y_i\alpha_i\sum_{m\in\mathbb{N}_M} q^{(\beta)}_{m\tau(x)\tau(i)}k_m(x_i,x)+b. \tag{8}$$

The resulting optimization algorithm for convex localized multiple kernel learning 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). Note that the proposed optimization approach can be potentially extended to an interleaved optimization strategy where the optimization of the MKL step is directly integrated into the SVM solver. It has been shown (Sonnenburg et al., 2006; Kloft et al., 2011) that such a strategy can increase the computational efficiency by up to 1-2 orders of magnitude (cf. Figure 7 in Kloft et al., 2011).

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

```
input: examples $\{(x_i, y_i)_{i=1}^n\} \subset \big(\mathcal{X} \times \{-1,1\}\big)^n$ together with cluster indices $\{\tau(i)\}_{i=1}^n$,
       $M$ base kernels $k_1, \dots, k_M$, and a positive semi-definite matrix $\Sigma^{-1}$.
1 initialize $\beta_{jm} = \sqrt[p]{1/M}$ for all $j \in \mathbb{N}_l, m \in \mathbb{N}_M$
2 while Optimality conditions are not satisfied do
3     calculate the kernel matrix $\tilde{k}$ by Eq. (3)
4     compute $\alpha$ by solving canonical SVM with $\tilde{k}$
5     compute $\|w_j^{(m)}\|_2^2$ for all $j, m$ by Eq. (7)
6     update $\beta_{jm}$ for all $j, m$ according to Eq. (6)
7 end
```

We remark that we also derive a complete dual problem removing the dependency on β_{jm}. Due to the lack of the space, we defer the detailed proof to the appendix:

Proposition 4 (CLMKL (COMPLETE) DUAL PROBLEM) *If Σ^{-1} is positive definite, then the completely dualized Lagrangian dual (dualized with respect to all variables) of Problem* (2) *becomes:*

$$\sup_{\substack{0 \le \alpha_i \le C \\ \sum_{i \in \mathbb{N}_n} \alpha_i y_i = 0}} \; \sup_{\substack{\gamma_{mj\tilde{j}} \\ m \in \mathbb{N}_M, j, \tilde{j} \in \mathbb{N}_l}} \Bigg\{ -\Bigg[\frac{1}{2} \sum_{j \in \mathbb{N}_l} \Big(\sum_{m \in \mathbb{N}_M} \Big\| \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i) - \sum_{i \in \mathbb{N}_n} \alpha_i y_i \gamma_{mj\tau(i)} \phi_m(x_i) \Big\|_2^{\frac{2p}{p-1}} \Big)^{\frac{p-1}{p}}$$

$$+ \frac{1}{2\mu} \sum_{m \in \mathbb{N}_M} \Big\| \big(\sum_{i \in \mathbb{N}_n} \alpha_i y_i \gamma_{mj\tau(i)} \phi_m(x_i) \big)_{j \in \mathbb{N}_l} \Big\|_{\Sigma}^2 \Bigg] + \sum_{i \in \mathbb{N}_n} \alpha_i \Bigg\}.$$

The above dual sheds further light onto the CLMKL optimization problem, and potentially can be exploited for the development of alternative optimization strategies that directly optimize the dual criterion (without the need of an two-step wrapper approach); such an approach has been taken in Sun et al. (2010) in the context of ℓ_p-norm MKL. Furthermore, solving the dual enables computing the duality gap, which can be used as a sound evaluation criterion for the optimization precision.

5. Rademacher complexity bounds

This section presents a theoretical analysis, showing, for the first time, that a localized approach to learning kernels can generalize to new and unseen data. In particular, we give a purely data-dependent bound on the generalization error. Our basic strategy is to plug the optimal β_{jm} established in Eq. (6) into the primal problem (2), thereby writing (2) as the following equivalent block-norm regularization problem:

$$\min_{w,b} \frac{1}{2}\sum_{j\in\mathbb{N}_l}\Big(\sum_{m\in\mathbb{N}_M}\|w_j^{(m)}\|_2^{\frac{2p}{p+1}}\Big)^{\frac{p+1}{p}} + \sum_{m\in\mathbb{N}_M}\frac{\mu}{2}\|w^{(m)}\|^2_{\Sigma^{-1}} \\ + C\sum_{i\in\mathbb{N}_n}\Big(1-y_i\sum_{m\in\mathbb{N}_M}\langle w^{(m)}_{\tau(i)},\phi_m(x_i)\rangle - y_i b\Big)_+. \quad (9)$$

Solving Eq. (9) amounts to performing empirical risk minimization in the hypothesis space

$$H_{p,\mu,D} := H_{p,\mu,D,M} = \Big\{ f_w : x \to \langle w_{\tau(x)}, \phi(x)\rangle : \sum_{j\in\mathbb{N}_l}\|w_j\|^2_{2,\frac{2p}{p+1}} + \mu\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|^2_{\Sigma^{-1}} \le D\Big\}.$$

The following theorem establishes a generalization error bound for CLMKL.

Theorem 5 (CLMKL GENERALIZATION ERROR BOUNDS) *Suppose that Σ^{-1} is positive definite and n is the sample size. Then, for any $0<\delta<1$ with probability at least $1-\delta$, the expected risk $\mathcal{E}(h) := \mathbb{E}[yh(x)\le 0]$ of any classifier $h \in H_{p,\mu,D}$ can be upper bounded by:*

$$\mathcal{E}(h) \le \mathcal{E}_{\mathbf{z}}(h) + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \\ \frac{2\sqrt{D}}{n}\inf_{\substack{0\le\theta\le 1\\ 2\le t\le \bar{p}^*}}\Big(\theta^2 t\sum_{j\in\mathbb{N}_l}\Big\|\Big(\sum_{i\in S_j}k_m(x_i,x_i)\Big)_{m=1}^M\Big\|_{\frac{t}{2}} + \frac{(1-\theta)^2}{\mu}\sum_{\substack{m\in\mathbb{N}_M\\ j\in\mathbb{N}_l}}\Sigma_{jj}\sum_{i\in S_j}k_m(x_i,x_i)\Big)^{1/2},$$

where $\mathcal{E}_{\mathbf{z}}(h) := \frac{1}{n}\sum_{i=1}^n(1-y_ih(x_i))_+$ is the empirical risk w.r.t. the hinge loss.

Remark 6 (Interpretation and Tightness) *The above error bound enjoys a mild dependence on the number of kernels. One can show (cf. Section C) that the dependence is $O(\log M)$ for $p \le (\log M - 1)^{-1}\log M$ and $O(M^{\frac{p-1}{2p}})$ otherwise, which recover the best known results for global MKL algorithms in Cortes et al. (2010); Kloft and Blanchard (2011); Kloft et al. (2011).*

Theorem 5 also suggests that the generalization performance of CLMKL is controlled by a weighted summation of the diagonal elements in the matrix Σ, with weights being proportional to the trace of the gram matrix on the associated clusters.

6. Empirical Studies

6.1 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). When the clusters $\{S_1,\ldots,S_l\}$ are not known in advance, they are computed through kernel k-means (e.g., Dhillon et al., 2004). To diminish k-mean's potential fluctuations due to random initialization of the cluster means, we repeat kernel k-means several times, and either select the one with the minimal clustering error (the summation of the squared distance between the examples and the associated nearest cluster) as the final partition, or train a single CLMKL model for each partition and then combine the resulting CLMKL models by performing *majority voting* on the binary predictions. We compare the performance attained by the proposed CLMKL to regular localized MKL (LMKL) (Gönen and Alpaydin, 2008), 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.

	CLMKL	LMKL	MKL	UNIF
$\sigma = 0.2$	98.3 ± 0.8	94.7 ± 1.4	94.8 ± 1.6	94.5 ± 1.6
$\sigma = 0.3$	91.4 ± 1.9	89.5 ± 1.8	89.2 ± 2.0	89.3 ± 1.7

Table 1: Performances achieved by LMKL, UNIF, MKL and the proposed CLMKL on the synthetic dataset. Here, σ is the standard deviation of the noise. The underlying parameter p is 1.

6.2 Controlled Experiments on Synthetic Data

We first experiment on a two-class synthetic dataset with positive and negative points lying on a disconnected hexagon with radius equal to 6 and 5, respectively, with additional corruptions by Gaussian noise with standard deviations σ. The figure to the right shows an example of such a synthetic dataset with 1000 examples and $\sigma = 0.2$. This dataset is interesting to us since the optimal combination of the features associated to the first and second coordinates vary along the six sides of the hexagon. We choose the linear kernels on the first and second coordinates as two base kernels for CLMKL, and apply k-means with 6 clusters to generate data partition. The correlation matrix Σ^{-1} is chosen as the graph Laplacian of an adjacency matrix W, where we set $W_{j\tilde{j}} = \exp(-\gamma d^2(S_j, S_{\tilde{j}}))$ with $d(S_j, S_{\tilde{j}})$ being the Euclidean distance between cluster S_j and $S_{\tilde{j}}$. The parameter γ is set as the reciprocal of the average distance among different clusters. We use one half of the dataset as the training set, and each half of the remaining as the validation set and test set. We tune the parameter C from the set $10^{\{-2,-1,\dots,2\}}$, and μ from the set $2^{\{2,4,6,8\}}$, based on the prediction accuracies on the validation set. For CLMKL and MKL, we simply set $p = 1$ in this experiment. For the baseline methods (LMKL, MKL, UNIF), we complement the linear features by adding the quadratic kernel $k(x, \tilde{x}) = \langle x, \tilde{x} \rangle^2$ as the third base kernel, which is a useful feature for this dataset since a circle (a quadratic function) with appropriate radius is expected to serve as a good predictor. Thus the addition of the quadratic kernel gives the baseline methods a potential advantage and serves as an additional sanity check of the robustness of the proposed algorithm.

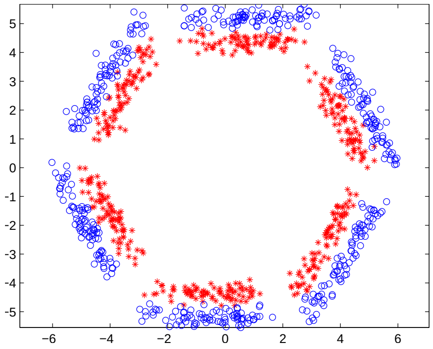


Table 1 shows the performance of the proposed CLMKL as well as the competitors. We can observe that the proposed CLMKL consistently achieves the best prediction accuracies with accuracy gains by up to 3.6%. Note that this improvement is achieved when the baseline methods are supplied with an additional quadratic kernel encoding valuable information for this synthetic data.

MKL	LMKL	CLMKL holdout	CLMKL Oracle	Li and Fei-Fei (2007)	Bo et al. (2011)	Liu et al. (2014)
90.23	87.36	90.80	91.38	73.8	85.7	89.95

Table 2: Prediction accuracies achieved by regular ℓ_p-norm MKL and the proposed CLMKL on the UIUC Sports Event dataset. The columns "Holdout" and "Oracle" show the prediction accuracies for the selected and optimal parameters, respectively. Liu et al. (2014) is the best known result from the literature.

	CLMKL, $l=4$		CLMKL, $l=8$		LMKL	MKL	UNIF
	holdout	Oracle	holdout	Oracle			
$p=1$	71.7 ± 0.4	72.8 ± 0.9	71.9 ± 0.4	73.7 ± 0.9	64.3	68.7	68.4
$p=1.14$	74.8 ± 0.4	75.2 ± 0.4	75.1 ± 0.5	75.4 ± 0.3		73.4	
$p=1.2$	74.9 ± 0.5	75.0 ± 0.6	74.7 ± 0.3	75.5 ± 0.6		74.2	
$p=1.33$	74.5 ± 0.4	75.0 ± 0.4	74.5 ± 0.4	74.7 ± 0.3		73.1	

Table 3: Prediction accuracies achieved by UNIF, LMKL, MKL and CLMKL on the protein folding class prediction task. The columns "Holdout" and "Oracle" show the prediction accuracies for the selected and optimal parameters, respectively. l indicates the number of clusters in CLMKL, and p indicates the type of regularizer on the kernel combination coefficients.

6.3 Visual Image Categorization—An Application from the Domain of Computer Vision

We experiment on the UIUC Sports event dataset (Li and Fei-Fei, 2007) consisting of 1574 images, each associated with one of 8 image classes (each class corresponding to a sport activity). We compute 12 bag-of-words features, each with a dictionary size of 512, resulting in 12 χ^2-Kernels (Zhang et al., 2007). The first 6 bag-of-words features are computed over SIFT features (Lowe, 2004) at three different scales and the two color channel sets RGB and opponent colors (van de Sande et al., 2010). The remaining 6 bag-of-words features are quantiles of color values at the same three scales and the same two color channel sets. For each channel within a set of color channels, the quantiles are concatenated. Local features are extracted at a grid of step size 5 on images that were down-scaled to 600 pixels in the largest dimension. Assignment of local features to visual words is done using rank-mapping (Binder et al., 2013). The kernel width of the kernels is set to the mean of the χ^2-distances. All kernels are multiplicatively normalized.

Following the setup of Liu et al. (2014) the dataset is split into 11 parts. One part is withheld to obtain the final performance measurements, and on the remaining 10 parts we perform 10-fold cross-validation for finding the optimal parameters. For CLMKL we employ kernel k-means with 3 clusters on the cross-validation parts. For CLMKL we apply majority voting over 8 separate clusterings, for each of which a separate predictor is trained for fixed parameters. The matrix Σ^{-1} is computed as $[(\exp(-\gamma d(S_j, S_{\tilde{j}})))_{j\tilde{j}}]^{-1}$ where as distance the χ^2-distances averaged over the cluster assignments are used. The two involved parameters γ and μ are determined by cross-validation.

We compare CLMKL to regular ℓ_p-norm MKL (Kloft et al., 2011), for which we employ a one-versus-all setup, running over ℓ_p-norms in $\{1.0625, 1.125, 1.333, 2\}$ and regularization constants in $\{10^{k/2}\}_{k=-2}^{k=5}$ (optima attained inside the respective grids). CLMKL uses the same set of ℓ_p-norms, regularization constants from $\{10^{k/2}\}_{k=0,\dots,5}$ and, due to time constraints, a subset of 18 combinations of the two parameters $(\gamma, \mu) \in \{10^{i/2}\}_{i=-4}^{i=0} \times \{2^i\}_{i=-4}^{i=4}$ is used to compute Σ^{-1}. Performance is measured through multi-class classification accuracy. Table 2 shows the results. The column "holdout" shows the prediction accuracy achieved by taking majority voting over predictors constructed based on different applications of kernel k-means with random initializations, while the column "Oracle" indicates the best prediction accuracy achieved by these models built on the output of kernel k-means with random initializations. We observe that CLMKL achieves a performance improvement by $0.5 - 1.2\%$ over the ℓ_p-norm MKL baseline. Comparing this to the best known results from the literature (Liu et al., 2014), we observe that this is, to our best knowledge, the highest result ever achieved on the UIUC dataset.

6.4 Protein Fold Prediction—An Application from the Domain of Computational Biology

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 reserved for testing. We use exactly the same 12 kernels used also in Campbell and Ying (2011); Kloft (2011); Kloft and Blanchard (2011) reflecting different features, e.g., van der Waals volume, polarity and hydrophobicity, relevant to the fold class prediction as base kernels. This is a non-sparse scenario for which Kloft (2011) achieved 74.4% accuracy using $\ell_{1.14}$-norm MKL.

To be in line with the previous experiments by Campbell and Ying (2011); Kloft (2011); Kloft and Blanchard (2011), we precisely replicate their experimental setup: we use the train/test split supplied by Campbell and Ying (2011) and perform CLMKL via one-versus-all strategy to tackle multiple classes. The correlation matrix Σ^{-1} is constructed in the same way as that in Section 6.2. The parameters are chosen by cross validation over $C \in 10^{\{-2,-1,\ldots,2\}}, \mu \in 2^{\{5,6,7\}}$. We consider ℓ_p-norm CLMKL models with $p \in \{1, 1.14, 1.2, 1.33\}$ and $l \in \{4, 8\}$. We repeat the experiment 10 times and report the mean prediction accuracies, as well as standard deviations in Table 3.

From the table, we observe that CLMKL has the potential to largely surpass its global counterpart ℓ_p-norm MKL. Note that we do not achieve the accuracy 74.4% for $\ell_{1.14}$-norm MKL reported in Kloft (2011), which is possibly due to different implementations of the ℓ_p-norm MKL algorithms. Nevertheless, CLMKL achieves accuracies more than 0.8% higher than the one reported in Kloft (2011), which is also higher than the one initially reported in Campbell and Ying (2011). For example, CLMKL with $l = 8, p = 1.14$ achieves an impressive accuracy of 75.1%.

7. Conclusion

We proposed a localized approach to learning kernels that admits generalization error bounds and can be phrased a convex optimization problem over a given or pre-obtained cluster structure. A key ingredient is the use of a graph-regularizer to couple the different local models. The theoretical analysis based on Rademacher complexity theory resulted in large deviation inequalities that connect the spectrum of the graph regularizer with the generalization capability of the learning algorithm. The proposed method is well suited for deployment in the domains of computer vision and computational biology: computational experiments showed that the proposed approach can achieve prediction accuracies higher than its global and non-convex local counterparts.

In future work, we will investigate alternative clustering strategies (including convex ones and soft clustering), and how to principally include the data partitioning into our framework, for instance, by constructing partitions that capture the local variation of prediction importance of different features, by solving the clustering step and the MKL optimization problem in a joint manner or by automatically learning the graph Laplacian using appropriate matrix regularization. Another research direction is to directly integrate the MKL step into the SVM solver, as pioneered by Sonnenburg et al. (2006). We expect that such an implementation would lead to a speed-up in computational efficiency by up to 1-2 orders of magnitude. We will also investigate extensions to other learning settings (Kloft et al., 2009; Mohri et al., 2015) and further applications (Kloft and Laskov, 2007; Nakajima et al., 2009; Binder et al., 2012; Kloft and Laskov, 2012; Kloft et al., 2014).

Acknowledgments

This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1.

References

Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

Peter Bartlett and Shahar 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. URL `http://svmcompbio.tuebingen.mpg.de`.

Alexander Binder, Shinichi Nakajima, Marius Kloft, Christina Müller, Wojciech Samek, Ulf Brefeld, Klaus-Robert Müller, and Motoaki Kawanabe. Insights from classifying visual concepts with multiple kernel learning. *PloS one*, 7(8):e38897, 2012.

Alexander Binder, Wojciech Samek, Klaus-Robert Müller, and Motoaki Kawanabe. Enhanced representation and multi-task learning for image annotation. *Computer Vision and Image Understanding*, 2013.

Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

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. ISBN 978-1-60558-516-1. Video `http://videolectures.net/icml09_cortes_clkh/`.

Corinna Cortes and Vladimir Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

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.

Manuel Fernández Delgado, Eva Cernadas, Senén Barro, and Dinani Gomes Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth 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.

Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

P.V. Gehler and S. Nowozin. Infinite kernel learning. In *Proceedings of the NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

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. URL `http://dl.acm.org/citation.cfm?id=1953048.2021071`.

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

Cijo Jose, Prasoon Goyal, Parv Aggrwal, and Manik Varma. Local deep kernel learning for efficient non-linear svm prediction. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 486–494, 2013.

Jean-Pierre Kahane. Some random series of functions, volume 5 of cambridge studies in advanced mathematics, 1985.

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

Marius Kloft and Gilles Blanchard. The local Rademacher complexity of ℓ_p-norm multiple kernel learning. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *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 and Pavel Laskov. A poisoning attack against online anomaly detection. *NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security*, 2007.

Marius Kloft and Pavel Laskov. Security analysis of online centroid anomaly detection. *The Journal of Machine Learning Research*, 13(1):3681–3724, 2012.

Marius Kloft, Shinichi Nakajima, and Ulf Brefeld. Feature selection for density level-sets. In *Machine Learning and Knowledge Discovery in Databases*, pages 692–704. Springer Berlin Heidelberg, 2009.

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

Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. Predicting mooc dropout over weeks using machine learning methods. *EMNLP 2014*, page 60, 2014.

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.

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.

Bao-Di Liu, Yu-Xiong Wang, Bin Shen, Yu-Jin Zhang, and Martial Hebert. Self-explanatory sparse representation for image classification. In *Computer Vision–ECCV 2014*, pages 600–616. Springer International Publishing, 2014.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. URL `http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94`.

Charles A Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, pages 1099–1125, 2005.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Mehryar Mohri, Afshin Rostamizadeh, and Dmitry Storcheus. Foundations of coupled nonlinear dimensionality reduction. *arXiv preprint arXiv:1509.08880v2*, 2015.

Shinichi Nakajima, Alexander Binder, Christina Müller, Wojciech Wojcikiewicz, Marius Kloft, Ulf Brefeld, Klaus-Robert Müller, and Motoaki Kawanabe. Multiple kernel learning for object classification. *Proceedings of the 12th Workshop on Information-based Induction Sciences*, 24, 2009.

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

Ryan M Rifkin and Ross A Lippert. Value regularization and fenchel duality. *The Journal of Machine Learning Research*, 8:441–479, 2007.

R Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1997.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Yan Song, Yan-Tao Zheng, Sheng Tang, Xiangdong Zhou, Yongdong Zhang, Shouxun Lin, and T-S Chua. Localized multiple kernel learning for realistic human action recognition in videos. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(9):1193–1202, 2011.

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

Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, 2010. URL `http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154`.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

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.

Yiming Ying and Colin Campbell. Generalization bounds for learning the kernel. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, COLT '09, Montreal, Quebec, Canada, 2009.

Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007. URL `http://dx.doi.org/10.1007/s11263-006-9794-4`.

Appendix A. Proofs on subproblems in Algorithm 1

A.1 Proof of Proposition 2

Proof of Proposition 2 The Lagrangian of the partial optimization problem w.r.t. $w_j^{(m)}$ and b is

$$L := \sum_{j\in\mathbb{N}_l}\sum_{m\in\mathbb{N}_M}\frac{\|w_j^{(m)}\|_2^2}{2\beta_{jm}} + \frac{\mu}{2}\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|_{\Sigma^{-1}}^2 + \\ \sum_{j\in\mathbb{N}_l}\sum_{i\in S_j}\alpha_i\Big(1-\xi_i-y_i\sum_{m\in\mathbb{N}_M}\langle w_j^{(m)},\phi_m(x_i)\rangle - y_i b\Big) + C\sum_{i\in\mathbb{N}_n}\xi_i - \sum_{i\in\mathbb{N}_n}v_i\xi_i, \quad \text{(A.1)}$$

where $\alpha_i \geq 0$ and $v_i \geq 0$ are the Lagrangian multipliers of the constraints.

Setting to zero the gradient of the Lagrangian w.r.t. the primal variables, we get

$$\frac{\partial L}{\partial w_j^{(m)}} = 0 \Rightarrow \frac{w_j^{(m)}}{\beta_{jm}} + \mu\sum_{\tilde{j}\in\mathbb{N}_l}\Sigma^{-1}_{j\tilde{j}}w_{\tilde{j}}^{(m)} - \sum_{i\in S_j}y_i\alpha_i\phi_m(x_i) = 0, \quad \text{(A.2)}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i\in\mathbb{N}_n}\alpha_i y_i = 0, \quad \text{(A.3)}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + v_i, \qquad \forall i \in \mathbb{N}_n. \quad \text{(A.4)}$$

Eq. (A.2) implies that

$$\sum_{j\in\mathbb{N}_l}\frac{||w_j^{(m)}||_2^2}{\beta_{jm}}+\mu||w^{(m)}||_{\Sigma^{-1}}^2=\sum_{j\in\mathbb{N}_l}\sum_{i\in S_j}\alpha_i y_i\langle w_j^{(m)},\phi_m(x_i)\rangle, \tag{A.5}$$

$$w_j^{(m)}=\sum_{\tilde{j}\in\mathbb{N}_l}q_{mj\tilde{j}}^{(\beta)}\sum_{i\in S_{\tilde{j}}}y_i\alpha_i\phi_m(x_i),\quad \forall j,m. \tag{A.6}$$

Plugging Eqs. (A.3), (A.4) into Eq. (A.1), the Lagrangian can be simplified as follows:

$$\begin{aligned}
L&=\sum_{i\in\mathbb{N}_n}\alpha_i+\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\frac{\|w_j^{(m)}\|_2^2}{2\beta_{jm}}+\sum_{m\in\mathbb{N}_M}\frac{\mu}{2}\|w^{(m)}\|_{\Sigma^{-1}}^2-\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\sum_{i\in S_j}\alpha_i y_i\langle w_j^{(m)},\phi_m(x_i)\rangle\\
&\overset{(A.5)}{=}\sum_{i\in\mathbb{N}_n}\alpha_i-\frac{1}{2}\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\langle w_j^{(m)},\sum_{i\in S_j}\alpha_i y_i\phi_m(x_i)\rangle\\
&\overset{(A.6)}{=}\sum_{i\in\mathbb{N}_n}\alpha_i-\frac{1}{2}\sum_{m\in\mathbb{N}_M}\sum_{j,\tilde{j}\in\mathbb{N}_l}q_{mj\tilde{j}}^{(\beta)}\sum_{i\in S_j,\tilde{i}\in S_{\tilde{j}}}y_i y_{\tilde{i}}\alpha_i\alpha_{\tilde{i}}\langle\phi_m(x_i),\phi_m(x_{\tilde{i}})\rangle\\
&=\sum_{i\in\mathbb{N}_n}\alpha_i-\frac{1}{2}\sum_{m\in\mathbb{N}_M}\sum_{i,\tilde{i}\in\mathbb{N}_n}y_i y_{\tilde{i}}\alpha_i\alpha_{\tilde{i}}q_{m\tau(i)\tau(\tilde{i})}^{(\beta)}k_m(x_i,x_{\tilde{i}})\\
&\overset{(3)}{=}\sum_{i\in\mathbb{N}_n}\alpha_i-\frac{1}{2}\sum_{i,\tilde{i}\in\mathbb{N}_n}y_i y_{\tilde{i}}\alpha_i\alpha_{\tilde{i}}\tilde{k}(x_i,x_{\tilde{i}}).
\end{aligned}$$

The proof is complete if we note the constraints established in Eqs. (A.3), (A.4). ■

A.2 Proof of Proposition 3

Proposition 3 contained in the main text gives a closed form solution for updating the kernel weights, a detailed proof of which is given in this appendix. Our discussion is largely based on the following lemma by Micchelli and Pontil (2005).

Lemma A.1 (Micchelli and Pontil, 2005, Lemma 26) *Let $a_i\geq 0, i\in\mathbb{N}_d$ and $1\leq r<\infty$. Then*

$$\min_{\eta:\eta_i\geq 0,\sum_{i\in\mathbb{N}_d}\eta_i^r\leq 1}\sum_{i\in\mathbb{N}_d}\frac{a_i}{\eta_i}=\Big(\sum_{i\in\mathbb{N}_d}a_i^{\frac{r}{r+1}}\Big)^{1+\frac{1}{r}}$$

and the minimum is attained at $\eta_i=a_i^{\frac{1}{r+1}}\Big(\sum_{k\in\mathbb{N}_d}a_k^{\frac{r}{r+1}}\Big)^{-\frac{1}{r}}$.

We are now ready to prove Proposition 3 as follows.

Proof of Proposition 3 Fixing the variables $w_j^{(m)}$ and b, the optimization problem (2) reduces to

$$\begin{aligned}
\min_{\beta}\quad&\frac{1}{2}\sum_{j\in\mathbb{N}_l,m\in\mathbb{N}_M}\beta_{jm}^{-1}\|w_j^{(m)}\|_2^2\\
\text{s.t.}\quad&\sum_{m\in\mathbb{N}_M}\beta_{jm}^p\leq 1,\forall j\in\mathbb{N}_l,\beta_{jm}\geq 0,\forall j\in\mathbb{N}_l,m\in\mathbb{N}_M.
\end{aligned}$$

This problem can be decomposed into l independent subproblems, one at each locality. For example, the subproblem at the j-th locality is as follows:

$$\begin{aligned}\min_{\beta} \quad & \frac{1}{2}\sum_{m\in\mathbb{N}_M} \beta_{jm}^{-1}\|w_j^{(m)}\|_2^2 \\ \text{s.t.} \quad & \sum_{m\in\mathbb{N}_M} \beta_{jm}^p \leq 1, \beta_{jm} \geq 0, \quad \forall m \in \mathbb{N}_M.\end{aligned}$$

Applying Lemma A.1 with $\alpha_m = \|w_j^{(m)}\|_2^2$, $\eta_m = \beta_{jm}$ and $r = p$ completes the proof. ■

Appendix B. Completely dualized problems

Proposition 2 gives a *partial* dual of the primal optimization problem (2). Alternatively, we derive here a *complete* dual problem removing the dependency on the kernel weights β_{jm}. This completes the analysis of the primal problem and can be potentially exploited (in future work) to access the duality gap of computed solutions or to derive an alternative optimization strategy (cf. Sun et al., 2010). We always assume that Σ is positive definite in this section. We consider a general loss function to give a unifying viewpoint, and our analysis is based on the notions of the Fenchel-Legendre conjugate (Boyd and Vandenberghe, 2004) and the infimal convolution (Rockafellar, 1997).

B.1 Lemmata used for complete dualization

For a function h, we denote by $h^*(x) = \sup_\mu[x^\top \mu - h(\mu)]$ its Fenchel-Legendre conjugate. The infimal convolution (short: Inf-convolution) of two functions f and g is defined by

$$(f \oplus g)(x) := \inf_y [f(x-y) + g(y)].$$

Lemma B.1 gives a relationship between the Fenchel-Legendre conjugate and the Inf-convolution.

Lemma B.1 (Rockafellar, 1997) *For any two functions f_1, f_2, we have $(f_1 + f_2)^*(x) = (f_1^* \oplus f_2^*)(x)$. Moreover, if f has a decomposable structure in the sense that $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$, i.e., f_1 and f_2 are functions defined on uncorrelated variables, then $(f_1 + f_2)^*(x) = (f_1^* + f_2^*)(x)$.*

For any norm $\|\cdot\|$, we denote by $\|\cdot\|_*$ its dual norm defined by $\|x\|_* = \sup_{\|\mu\|=1}\langle x, \mu\rangle$. The Fenchel-Legendre conjugate of square norm takes the following form (Rockafellar, 1997):

$$(\frac{1}{2}\|\cdot\|^2)^* = \frac{1}{2}\|\cdot\|_*^2. \tag{B.1}$$

Lemma B.2 establishes the dual norm for a Σ-norm. The result is well-known if $\mathcal{H}$ is the 1-dimensional Euclidean space.

Lemma B.2 *Let $\mathcal{H}$ be a Hilbert space and Σ be a $l \times l$ positive definite matrix. The dual norm of the Σ-norm defined by $\|(w_1, \ldots, w_l)\|_\Sigma = \Big(\sum_{j,\tilde{j}\in\mathbb{N}_l} \Sigma_{j\tilde{j}}\langle w_j, w_{\tilde{j}}\rangle\Big)^{1/2}$ is the Σ^{-1}-norm.*

Proof For any two elements $w=(w_1,\ldots,w_l), v=(v_1,\ldots,v_l)\in\underbrace{\mathcal{H}\times\cdots\times\mathcal{H}}_{l}$, we first establish the following inequality:

$$\langle(v_1,\ldots,v_l),(w_1,\ldots,w_l)\rangle\leq\|(v_1,\ldots,v_l)\|_{\Sigma^{-1}}\|(w_1,\ldots,w_l)\|_{\Sigma}. \tag{B.2}$$

Let $\mu_1,\ldots,\mu_l\in\mathbb{R}^l$ be the eigenvectors of Σ with $\lambda_1,\ldots,\lambda_l$ being the corresponding eigenvalues. According to the single value decomposition, we have

$$\Sigma=\sum_{k\in\mathbb{N}_l}\lambda_k\mu_k\mu_k^\top,\quad\Sigma^{-1}=\sum_{k\in\mathbb{N}_l}\lambda_k^{-1}\mu_k\mu_k^\top,$$

from which we know

$$\|w\|_\Sigma^2=\sum_{k\in\mathbb{N}_l}\lambda_k\|w\|^2_{\mu_k\mu_k^\top}=\sum_{k\in\mathbb{N}_l}\lambda_k\langle\sum_{j\in\mathbb{N}_l}\mu_{kj}w_j,\sum_{j\in\mathbb{N}_l}\mu_{kj}w_j\rangle.$$

Therefore,

$$\begin{aligned}\|(w_1,\ldots,w_l)\|_\Sigma\|(v_1,\ldots,v_l)\|_{\Sigma^{-1}}&=\Big(\sum_{k\in\mathbb{N}_l}\lambda_k\|\sum_{j\in\mathbb{N}_l}\mu_{kj}w_j\|_2^2\Big)^{1/2}\Big(\sum_{k\in\mathbb{N}_l}\lambda_k^{-1}\|\sum_{j\in\mathbb{N}_l}\mu_{kj}v_j\|_2^2\Big)^{1/2}\\&\overset{\text{C. S.}}{\geq}\sum_{k\in\mathbb{N}_l}\|\sum_{j\in\mathbb{N}_l}\mu_{kj}w_j\|_2\|\sum_{j\in\mathbb{N}_l}\mu_{kj}v_j\|_2\\&\geq\sum_{k\in\mathbb{N}_l}\langle\sum_{j\in\mathbb{N}_l}\mu_{kj}w_j,\sum_{j\in\mathbb{N}_l}\mu_{kj}v_j\rangle\\&=\sum_{j,\tilde{j}\in\mathbb{N}_l}\langle w_j,v_{\tilde{j}}\rangle\Big(\sum_{k\in\mathbb{N}_l}\mu_{kj}\mu_{k\tilde{j}}\Big).\end{aligned} \tag{B.3}$$

Since $\sum_{k\in\mathbb{N}_l}\mu_k\mu_k^\top$ is the identity matrix, we know that $\sum_{k\in\mathbb{N}_l}\mu_{kj}\mu_{k\tilde{j}}=\delta_{j\tilde{j}}$. Plugging this identity into the above inequality yields Eq. (B.2).

Next, we need to show that for any $w=(w_1,\ldots,w_l)$ there exists an $v=(v_1,\ldots,v_l)$ for which Eq. (B.2) holds as an equality. Introduce the invertible matrix $B=\big(\frac{1}{\lambda_1}\mu_1,\ldots,\frac{1}{\lambda_l}\mu_l\big)^\top$ and denote by B^{-1} its inverse. Then, we have

$$\frac{1}{\lambda_k}\sum_{j\in\mathbb{N}_l}\mu_{kj}B^{-1}_{j\tilde{k}}=\delta_{k\tilde{k}}. \tag{B.4}$$

Introduce

$$v_k:=\sum_{j\in\mathbb{N}_l}B^{-1}_{kj}\Big(\sum_{\tilde{j}\in\mathbb{N}_l}\mu_{j\tilde{j}}w_{\tilde{j}}\Big),\quad\forall k\in\mathbb{N}_l.$$

Then, it follows from Eq. (B.4) that

$$\begin{aligned}\sum_{j\in\mathbb{N}_l}\frac{1}{\lambda_k}\mu_{kj}v_j&=\sum_{j\in\mathbb{N}_l}\frac{1}{\lambda_k}\mu_{kj}\Big(\sum_{\tilde{k}\in\mathbb{N}_l}B^{-1}_{j\tilde{k}}\Big(\sum_{\tilde{j}\in\mathbb{N}_l}\mu_{\tilde{k}\tilde{j}}w_{\tilde{j}}\Big)\Big)=\sum_{\tilde{k}\in\mathbb{N}_l}\Big(\sum_{j\in\mathbb{N}_l}\frac{1}{\lambda_k}\mu_{kj}B^{-1}_{j\tilde{k}}\Big)\Big(\sum_{\tilde{j}\in\mathbb{N}_l}\mu_{\tilde{k}\tilde{j}}w_{\tilde{j}}\Big)\\&\overset{\text{(B.4)}}{=}\sum_{\tilde{k}\in\mathbb{N}_l}\delta_{k\tilde{k}}\Big(\sum_{\tilde{j}\in\mathbb{N}_l}\mu_{\tilde{k}\tilde{j}}w_{\tilde{j}}\Big)=\sum_{\tilde{j}\in\mathbb{N}_l}\mu_{k\tilde{j}}w_{\tilde{j}}.\end{aligned}$$

For any w, v satisfying the above relation, we have

$$\lambda_k^2 \| \sum_{j \in \mathbb{N}_l} \mu_{kj} w_j \|_2^2 = \| \sum_{j \in \mathbb{N}_l} \mu_{kj} v_j \|_2^2, \quad \| \sum_{j \in \mathbb{N}_l} \mu_{kj} w_j \|_2 \| \sum_{j \in \mathbb{N}_l} \mu_{kj} v_j \|_2 = \langle \sum_{j \in \mathbb{N}_l} \mu_{kj} w_j, \sum_{j \in \mathbb{N}_l} \mu_{kj} v_j \rangle,$$

and therefore the inequality (B.3) holds indeed as an equality. The proof is complete. ■

B.2 Proofs on complete dualization problems

The convex localized MKL model given in (2) can be extended to a general convex loss function:

$$\begin{aligned} \min_{w, t_i, \beta, b} \ & \sum_{j \in \mathbb{N}_l, m \in \mathbb{N}_M} \frac{\|w_j^{(m)}\|_2^2}{2\beta_{jm}} + \frac{\mu}{2} \sum_{m \in \mathbb{N}_M} \|w^{(m)}\|_{\Sigma^{-1}}^2 + C \sum_{i \in \mathbb{N}_n} \ell(t_i, y_i) \\ \text{s.t.} \ & \sum_{m \in \mathbb{N}_M} \beta_{jm}^p \le 1, \ \forall j \in \mathbb{N}_l, \beta_{jm} \ge 0, \ \forall j \in \mathbb{N}_l, m \in \mathbb{N}_M \\ & \sum_{m \in \mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x_i) \rangle + b = t_i, \ \forall i \in S_j, j \in \mathbb{N}_l. \end{aligned} \tag{B.5}$$

Here $\ell(t_i, y_i)$ is a general loss function measuring the error incurred from using t_i to predict y_i. The following theorem gives the complete dual problem for the above convex localized MKL.

Problem B.3 (COMPLETELY DUALIZED DUAL PROBLEM FOR GENERAL LOSS FUNCTIONS) *Let $\ell(t, y) : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex function w.r.t. t for any y. Assume that Σ^{-1} is positive definite. Then we have the following complete dual problem for the formulation* (B.5)*:*

$$\sup_{\sum_{i \in \mathbb{N}_n} \alpha_i = 0} \Bigg\{ -C \sum_{i \in \mathbb{N}_n} \ell^*(-\frac{\alpha_i}{C}, y_i) - \\ \Bigg[\Bigg[\frac{1}{2} \sum_{j \in \mathbb{N}_l} \Big(\sum_{m \in \mathbb{N}_M} \| \sum_{i \in S_j} \alpha_i \phi_m(x_i) \|_2^{\frac{2p}{p-1}} \Big)^{\frac{p-1}{p}} \Bigg] \oplus \Bigg[\frac{1}{2\mu} \sum_{m \in \mathbb{N}_M} \| \big(\sum_{i \in S_j} \alpha_i \phi_m(x_i) \big)_{j \in \mathbb{N}_l} \|_{\Sigma}^2 \Bigg]\Bigg] \Bigg\}.$$

Proof Using Proposition 3 to get the optimal β_{jm}, the problem (B.5) is equivalent to

$$\begin{aligned} \inf_{w, b, t_i} \ & \frac{1}{2} \sum_{j \in \mathbb{N}_l} \left(\sum_{m \in \mathbb{N}_M} \|w_j^{(m)}\|_2^{\frac{2p}{p+1}} \right)^{\frac{p+1}{p}} + \frac{\mu}{2} \sum_{m \in \mathbb{N}_M} \|w^{(m)}\|_{\Sigma^{-1}}^2 + C \sum_{i \in \mathbb{N}_n} \ell(t_i, y_i) \\ \text{s.t.} \ & \sum_{m \in \mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x_i) \rangle + b = t_i, \ \forall i \in S_j, j \in \mathbb{N}_l. \end{aligned}$$

According to the definition of Fenchel-Legendre conjugate and its relationship to Inf-convolution established in Lemma B.1, the Lagrangian saddle point problem translates to

$$\begin{aligned}
&\sup_{\alpha_i} \inf_{w,b,t} \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}} + \frac{\mu}{2} \sum_{m\in\mathbb{N}_M} \|w^{(m)}\|_{\Sigma^{-1}}^2 + C \sum_{i\in\mathbb{N}_n} \ell(t_i, y_i) \\
&\qquad\qquad - \sum_{j\in\mathbb{N}_l} \sum_{i\in S_j} \alpha_i \big(\sum_{m\in\mathbb{N}_M} \langle w_j^{(m)}, \phi_m(x_i) \rangle + b - t_i \big) \\
&= \sup_{\alpha_i} \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 \\
&\quad - \sup_w \Big[\sum_{j\in\mathbb{N}_l} \sum_{m\in\mathbb{N}_M} \langle w_j^{(m)}, \sum_{i\in S_j} \alpha_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}} - \frac{\mu}{2} \sum_{m\in\mathbb{N}_M} \|w^{(m)}\|_{\Sigma^{-1}}^2 \Big] \Big\} \\
&\overset{\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) \\
&\qquad - \Big[\frac{1}{2} \sum_{j\in\mathbb{N}_l} \Big(\sum_{m\in\mathbb{N}_M} \| \sum_{i\in S_j} \alpha_i \phi_m(x_i) \|_2^{\frac{2p}{p+1}} \Big)^{\frac{p+1}{p}} + \frac{\mu}{2} \sum_{m\in\mathbb{N}_M} \| \big(\sum_{i\in S_j} \alpha_i \phi_m(x_i) \big)_{j\in\mathbb{N}_l} \|_{\Sigma^{-1}}^2 \Big]^* \Big\} \\
&\overset{\text{Lem. } B.1}{=} \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) \\
&\quad - \Big[\Big[\frac{1}{2} \sum_{j\in\mathbb{N}_l} \Big(\sum_{m\in\mathbb{N}_M} \| \sum_{i\in S_j} \alpha_i \phi_m(x_i) \|_2^{\frac{2p}{p+1}} \Big)^{\frac{p+1}{p}} \Big]^* \oplus \Big[\frac{\mu}{2} \sum_{m\in\mathbb{N}_M} \| \big(\sum_{i\in S_j} \alpha_i \phi_m(x_i) \big)_{j\in\mathbb{N}_l} \|_{\Sigma^{-1}}^2 \Big]^* \Big] \Big\} \\
&\overset{\text{Lem. } B.1}{=} \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) \\
&\quad - \Big[\Big[\sum_{j\in\mathbb{N}_l} \Big(\frac{1}{2} \Big\| \Big(\| \sum_{i\in S_j} \alpha_i \phi_m(x_i) \|_2 \Big)_{m\in\mathbb{N}_M} \Big\|_{\frac{2p}{p+1}}^2 \Big)^* \Big] \oplus \Big[\sum_{m\in\mathbb{N}_M} \Big(\frac{\mu}{2} \| \big(\sum_{i\in S_j} \alpha_i \phi_m(x_i) \big)_{j\in\mathbb{N}_l} \|_{\Sigma^{-1}}^2 \Big)^* \Big] \Big] \Big\} \\
&\overset{\text{(B.1)}}{=} \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) - \\
&\qquad \Big[\Big[\frac{1}{2} \sum_{j\in\mathbb{N}_l} \Big(\sum_{m\in\mathbb{N}_M} \| \sum_{i\in S_j} \alpha_i \phi_m(x_i) \|_2^{\frac{2p}{p-1}} \Big)^{\frac{p-1}{p}} \Big] \oplus \Big[\frac{1}{2\mu} \sum_{m\in\mathbb{N}_M} \| \big(\sum_{i\in S_j} \alpha_i \phi_m(x_i) \big)_{j\in\mathbb{N}_l} \|_{\Sigma}^2 \Big] \Big] \Big\}.
\end{aligned}$$

In the last step of the above deduction, we have used the fact that Σ^{-1}-norm and Σ-norm, ℓ_p-norm and $\ell_{\frac{p}{p-1}}$-norm are two dual-norm pairs. ■

We can now prove the complete dual problem established in Proposition 4 by plugging the Fenchel conjugate function of the hinge loss into Problem B.3.

Proof of Proposition 4 Note that the Fenchel-Legendre conjugate of the hinge loss is $\ell^*(t, y) = \frac{t}{y}$ (a function of t) if $-1 \leq \frac{t}{y} \leq 0$ and ∞ elsewise (Rifkin and Lippert, 2007). Recall the identity $(\eta f)^*(x) = \eta f^*(x/\eta)$. Hence, for each i, the term $\ell^*(-\frac{\alpha_i}{C}, y_i)$ translates to $-\frac{\alpha_i}{C y_i}$, provided that

$0 \leq \frac{\alpha_i}{y_i} \leq C$. With a variable substitution of the form $\alpha_i^{\text{new}} = \frac{\alpha_i}{y_i}$, the complete dual problem established in Theorem B.3 now becomes

$$\sup_{\substack{0 \leq \alpha_i \leq C \\ \sum_{i \in \mathbb{N}_n} \alpha_i y_i = 0}} \sum_{i \in \mathbb{N}_n} \alpha_i -$$

$$\Big[\frac{1}{2\mu} \sum_{m \in \mathbb{N}_M} \|\big(\sum_{i \in S_j} \alpha_i y_i \phi_m(x_i)\big)_{j \in \mathbb{N}_l}\|_\Sigma^2 \oplus \Big(\frac{1}{2} \sum_{j \in \mathbb{N}_l} \Big(\sum_{m \in \mathbb{N}_M} \| \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i) \|_2^{\frac{2p}{p-1}} \Big)^{\frac{p-1}{p}} \Big)\Big]$$

$$\overset{\text{Def}}{=} \sup_{\substack{0 \leq \alpha_i \leq C \\ \sum_{i \in \mathbb{N}_n} \alpha_i y_i = 0}} \sup_{\theta_j^{(m)}} \Big\{ \sum_{i \in \mathbb{N}_n} \alpha_i - \Big[\frac{1}{2\mu} \sum_{m \in \mathbb{N}_M} \|\big(\theta_j^{(m)}\big)_{j \in \mathbb{N}_l}\|_\Sigma^2$$

$$+ \Big(\frac{1}{2} \sum_{j \in \mathbb{N}_l} \Big(\sum_{m \in \mathbb{N}_M} \| \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i) - \theta_j^{(m)} \|_2^{\frac{2p}{p-1}} \Big)^{\frac{p-1}{p}} \Big)\Big]\Big\}.$$

The optimal $\theta_j^{(m)}$ satisfies the following K.K.T. condition

$$\Big(\sum_{m \in \mathbb{N}_M} \Big\| \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i) - \theta_j^{(m)} \Big\|_2^{\frac{2p}{p-1}} \Big)^{\frac{-1}{p}} \Big\| \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i) - \theta_j^{(m)} \Big\|_2^{\frac{2}{p-1}} \big(\theta_j^{(m)} - \sum_{i \in S_j} \alpha_i y_i \phi_m(x_i)\big)$$
$$= -\frac{1}{\mu} \sum_{\tilde{j} \in \mathbb{N}_l} \Sigma_{j\tilde{j}} \theta_{\tilde{j}}^{(m)}.$$

Solving the above equation shows that the optimal $\theta_j^{(m)}$ takes the following form

$$\theta_j^{(m)} = \sum_{i \in \mathbb{N}_n} \alpha_i y_i \gamma_{mj\tau(i)} \phi_m(x_i), \qquad \forall j \in \mathbb{N}_l, m \in \mathbb{N}_M.$$

Plugging the above identity back into the Langangian saddle point problem, we derive the complete dual problem in the proposition. ■

Appendix C. Proof of Generalization Error Bounds (Theorem 5)

This section presents the proof for the generalization error bounds provided in Section 5. Our basic tool is the data-dependent complexity measure called the Rademacher complexity (Bartlett and Mendelson, 2002).

Definition C.1 (Rademacher complexity) 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 $\{\pm 1\}$-valued random variables.

The following theorem establishes the Rademacher complexity bounds for CLMKL machines. Denote $\bar{p} = \frac{2p}{p+1}$ for any $p \geq 1$ and observe that $\bar{p} \leq 2$, which implies $\bar{p}^* \geq 2$.

Theorem C.2 (CLMKL RADEMACHER COMPLEXITY BOUNDS) *If Σ^{-1} is positive definite, then the empirical Rademacher complexity of $H_{p,\mu,D}$ can be controlled by*

$$\hat{R}_n(H_{p,\mu,D}) \leq \frac{\sqrt{D}}{n} \inf_{\substack{0 \leq \theta \leq 1 \\ 2 \leq t \leq \bar{p}^*}} \left(\theta^2 t \sum_{j \in \mathbb{N}_l} \Big\| \Big(\sum_{i \in S_j} k_m(x_i, x_i) \Big)_{m=1}^{M} \Big\|_{\frac{t}{2}} + \frac{(1-\theta)^2}{\mu} \sum_{\substack{m \in \mathbb{N}_M \\ j \in \mathbb{N}_l}} \Sigma_{jj} \sum_{i \in S_j} k_m(x_i, x_i) \right)^{1/2}.$$

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

$$\hat{R}_n(H_{p,\mu,D}) \leq \sqrt{\frac{DB}{n}} \inf_{\substack{0 \leq \theta \leq 1 \\ 2 \leq t \leq \bar{p}^*}} \left(\theta^2 t M^{\frac{2}{t}} + \frac{(1-\theta)^2}{\mu} M \max_{j \in \mathbb{N}_l} \Sigma_{jj} \right)^{1/2}.$$

Tightness of the bound It can be checked that the function $x \to xM^{2/x}$ is decreasing along the interval $(0, 2\log M)$ and increasing along the interval $(2\log M, \infty)$. Therefore, under the boundedness assumption $k_m(x,x) \leq B$ the Rademacher complexity can be further controlled by

$$\hat{R}_n(H_{p,\mu,D}) \leq \sqrt{\frac{DB}{n}} \times \begin{cases} \min\Big(\big(2e\log M\big)^{\frac{1}{2}}, \big(M\mu^{-1} \max_{j \in \mathbb{N}_l} \Sigma_{jj}\big)^{\frac{1}{2}} \Big), & \text{if } p \leq \frac{\log M}{\log M - 1}, \\ \min\Big(\big(\frac{2p}{p-1}\big)^{\frac{1}{2}} M^{\frac{p-1}{2p}}, \big(M\mu^{-1} \max_{j \in \mathbb{N}_l} \Sigma_{jj}\big)^{\frac{1}{2}} \Big), & \text{otherwise,} \end{cases}$$

from which it is clear that our Rademacher complexity bounds enjoy a mild dependence on the number of kernels. The dependence is $O(\log M)$ for $p \leq (\log M - 1)^{-1} \log M$ and $O(M^{\frac{p-1}{2p}})$ otherwise. 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 proof of Theorem C.2 is based on the following lemmata.

Lemma C.3 (Khintchine-Kahane inequality (Kahane, 1985)) *Let $v_1, \ldots, v_n \in \mathcal{H}$. Then, for any $q \geq 1$, it holds*

$$\mathbb{E}_{\boldsymbol{\sigma}} \Big\| \sum_{i \in \mathbb{N}_n} \sigma_i v_i \Big\|_2^q \leq \Big(q \sum_{i \in \mathbb{N}_n} \|v_i\|_2^2 \Big)^{\frac{q}{2}}.$$

Lemma C.4 (Block-structured Hölder inequality (Kloft and Blanchard, 2012)) *Let*

$$x = (x^{(1)}, \ldots, x^{(n)}), y = (y^{(1)}, \ldots, y^{(n)}) \in \mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_n.$$

Then, for any $p \geq 1$, it holds $\langle x, y \rangle \leq \|x\|_{2,p} \|y\|_{2,p^*}$.

Proof of Theorem C.2 Firstly, for any $\bar{t} \geq 1$ we can apply a block-structured version of Hölder inequality to bound $\sum_{i \in \mathbb{N}_n} \sigma_i f_w(x_i)$ by

$$\begin{aligned} \sum_{i \in \mathbb{N}_n} \sigma_i f_w(x_i) &= \sum_{i \in \mathbb{N}_n} \sigma_i \langle w_{\tau(i)}, \phi(x_i) \rangle = \sum_{j \in \mathbb{N}_l} \sum_{i \in S_j} \sigma_i \langle w_j, \phi(x_i) \rangle \\ &= \sum_{j \in \mathbb{N}_l} \Big\langle w_j, \sum_{i \in S_j} \sigma_i \phi(x_i) \Big\rangle \overset{\text{Hölder}}{\leq} \sum_{j \in \mathbb{N}_l} \|w_j\|_{2,\bar{t}} \Big\| \sum_{i \in S_j} \sigma_i \phi(x_i) \Big\|_{2,\bar{t}^*}. \end{aligned}$$

Alternatively, we can also control $\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)$ by

$$\begin{aligned}\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)&=\sum_{j\in\mathbb{N}_l}\big\langle w_j,\sum_{i\in S_j}\sigma_i\phi(x_i)\big\rangle=\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\Big\langle w_j^{(m)},\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big\rangle\\&=\sum_{m\in\mathbb{N}_M}\Big\langle w^{(m)},\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\rangle\\&\leq\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|_{\Sigma^{-1}}\Big\|\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\|_{\Sigma},\end{aligned}$$

where in the last step of the above deduction we have used the fact that Σ-norm is the dual norm of Σ^{-1}-norm (Lemma B.2).

Combining the above two inequalities together and using the trivial identity $\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)=\theta\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)+(1-\theta)\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)$, for any $0\leq\theta\leq1$ and any $t\geq1$ we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f_w\in H_{t,\mu,D}}\sum_{i\in\mathbb{N}_n}\sigma_i f_w(x_i)\\&\leq\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f_w\in H_{t,\mu,D}}\Big[\theta\sum_{j\in\mathbb{N}_l}\|w_j\|_{2,\bar{t}}\Big\|\sum_{i\in S_j}\sigma_i\phi(x_i)\Big\|_{2,\bar{t}^*}+(1-\theta)\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|_{\Sigma^{-1}}\Big\|\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\|_{\Sigma}\Big]\\&\overset{\text{C.-S.}}{\leq}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f_w\in H_{t,\mu,D}}\Big(\sum_{j\in\mathbb{N}_l}\|w_j\|_{2,\bar{t}}^2+\mu\sum_{m\in\mathbb{N}_M}\|w^{(m)}\|_{\Sigma^{-1}}^2\Big)^{1/2}\\&\qquad\times\Big(\theta^2\sum_{j\in\mathbb{N}_l}\Big\|\sum_{i\in S_j}\sigma_i\phi(x_i)\Big\|_{2,\bar{t}^*}^2+\frac{(1-\theta)^2}{\mu}\sum_{m\in\mathbb{N}_M}\Big\|\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\|_{\Sigma}^2\Big)^{1/2}\\&\overset{\text{Jensen}}{\leq}\Big(D\mathbb{E}_{\boldsymbol{\sigma}}\Big[\theta^2\sum_{j\in\mathbb{N}_l}\Big\|\sum_{i\in S_j}\sigma_i\phi(x_i)\Big\|_{2,\bar{t}^*}^2+\frac{(1-\theta)^2}{\mu}\sum_{m\in\mathbb{N}_M}\Big\|\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\|_{\Sigma}^2\Big]\Big)^{1/2}.\end{aligned}\tag{C.1}$$

For any $j\in\mathbb{N}_l$, the Khintchine-Kahane (K.-K.) inequality and Jensen inequality (since $\bar{t}^*\geq2$) permit us to bound $\mathbb{E}_{\boldsymbol{\sigma}}\Big\|\sum_{i\in S_j}\sigma_i\phi(x_i)\Big\|_{2,\bar{t}^*}^2$ by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\Big\|\sum_{i\in S_j}\sigma_i\phi(x_i)\Big\|_{2,\bar{t}^*}^2&\overset{\text{Def.}}{=}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sum_{m\in\mathbb{N}_M}\Big\|\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big\|_2^{\bar{t}^*}\Big]^{\frac{2}{\bar{t}^*}}\overset{\text{Jensen}}{\leq}\Big[\mathbb{E}_{\boldsymbol{\sigma}}\sum_{m\in\mathbb{N}_M}\Big\|\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big\|_2^{\bar{t}^*}\Big]^{\frac{2}{\bar{t}^*}}\\&\overset{\text{K.-K.}}{\leq}\Big[\sum_{m\in\mathbb{N}_M}\Big(\bar{t}^*\sum_{i\in S_j}\|\phi_m(x_i)\|_2^2\Big)^{\frac{\bar{t}^*}{2}}\Big]^{\frac{2}{\bar{t}^*}}=\bar{t}^*\Big[\sum_{m\in\mathbb{N}_M}\Big(\sum_{i\in S_j}k_m(x_i,x_i)\Big)^{\frac{\bar{t}^*}{2}}\Big]^{\frac{2}{\bar{t}^*}}\\&=\bar{t}^*\Big\|\Big(\sum_{i\in S_j}k_m(x_i,x_i)\Big)_{m=1}^{M}\Big\|_{\frac{\bar{t}^*}{2}}.\end{aligned}$$

For any $m \in \mathbb{N}_M$, we also have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\Big\|\Big(\sum_{i\in S_j}\sigma_i\phi_m(x_i)\Big)_{j=1}^{l}\Big\|_{\Sigma}^{2} &= \sum_{j,\tilde{j}\in\mathbb{N}_l}\Sigma_{j\tilde{j}}\mathbb{E}_{\boldsymbol{\sigma}}\Big\langle\sum_{i\in S_j}\sigma_i\phi_m(x_i),\sum_{\tilde{i}\in S_{\tilde{j}}}\sigma_{\tilde{i}}\phi_m(x_{\tilde{i}})\Big\rangle\\ &= \sum_{j\in\mathbb{N}_l}\mathbb{E}_{\boldsymbol{\sigma}}\Sigma_{jj}\Big\langle\sum_{i\in S_j}\sigma_i\phi_m(x_i),\sum_{\tilde{i}\in S_j}\sigma_{\tilde{i}}\phi_m(x_{\tilde{i}})\Big\rangle\\ &= \sum_{j\in\mathbb{N}_l}\Sigma_{jj}\sum_{i\in S_j}k_m(x_i,x_i).\end{aligned}$$

Plugging the above two inequalities into Eq. (C.1) and noticing the trivial inequality $\|w_j\|_{2,\bar{t}} \leq \|w_j\|_{2,\bar{p}}, \forall t \geq p \geq 1$, we get the following bound for any $0 \leq \theta \leq 1$:

$$\begin{aligned}&\hat{R}_n(H_{p,\mu,D}) \leq \inf_{t\geq p}\hat{R}_n(H_{t,\mu,D})\\ &\leq \frac{\sqrt{D}}{n}\inf_{t\geq p}\Bigg(\theta^2\bar{t}^*\sum_{j\in\mathbb{N}_l}\Big\|\Big(\sum_{i\in S_j}k_m(x_i,x_i)\Big)_{m=1}^{M}\Big\|_{\frac{\bar{t}^*}{2}}+\frac{(1-\theta)^2}{\mu}\sum_{m\in\mathbb{N}_M}\sum_{j\in\mathbb{N}_l}\Sigma_{jj}\sum_{i\in S_j}k_m(x_i,x_i)\Bigg)^{1/2}.\end{aligned}$$

The above inequality can be equivalently written as the first inequality of the theorem. The second inequality follows directly from the boundedness assumption and the fact that

$$\sum_{j\in\mathbb{N}_l}\Sigma_{jj}|S_j| \leq \max_{j\in\mathbb{N}_l}\Sigma_{jj}n.$$

■

Proof of Theorem 5 The proof now simply follows by plugging in the bound of Theorem C.2 into Theorem 7 of Bartlett and Mendelson (2002). ■

Appendix D. Parameter sets for the CLMKL on the UIUC Sports event dataset

We have chosen the following pairs of the two parameters (μ, γ): $(0.0612, 0.0100)$, $(0.1250, 0.0100)$, $(0.2500, 0.0100)$, $(0.0612, 0.0316)$, $(0.1250, 0.0316)$, $(0.0612, 0.1000)$, $(0.1250, 0.1000)$,$(2.0000, 0.1000)$, $(0.0612, 0.3162)$, $(0.2500, 0.3162)$, $(1.0000, 0.3162)$, $(16.0000, 0.3162)$, $(0.0612, 1.0000)$, $(0.1250, 1.0000)$, $(0.2500, 1.0000)$, $(0.5000, 1.0000)$, $(1.0000, 1.0000)$, $(2.0000, 1.0000)$, $(8.0000, 1.0000)$. The parameters as selected by 10-fold crossvalidation for CLMKL were: $\ell_p = 1.333$,$\mu = 2.0$,$\gamma = 1.0$