A new algorithm to learn an optimal kernel based on Fenchel duality

S. Mosci, M. Santoro, A. Verri, and S. Villa
DISI, Università di Genova, Italy

L. Rosasco
Center for Biological and Computational Learning, Massachusetts Institute of Technology
DISI, Università di Genova, Italy

1 Introduction and Problem Statement

Kernel methods have strongly influenced the state of the art of supervised learning over the past decades. Successful applications of kernel-based algorithms are widespread, and the interest on new variants of the basic methods still keep growing. Since a key property of a supervised learning algorithm is to achieve good generalization along with good performance, an important issue is to reduce as much as possible the need for human intervention throughout the learning process. That means, for instance, one should reduce to the bare essentials any *subjective* choice during the selection of the underlying statistical model and the tuning of the parameters. A great deal of efforts have been made to address the problem of automatic parameter selection within the field of regularized regression and classification. However, only in the last few years specific attention has been payed to the automatic selection of an *appropriate* hypothesis space, i.e. the problem of choosing the optimal kernel, given a specific set of input-output pairs (e.g., see the seminal contributions [6] and [3]).

In this paper, starting from the formulation of the problem due to the authors of [7] and [1], we propose a new algorithm to select the optimal combination of a prescribed collection of basic kernels for some specific supervised task. More specifically, let's denote $\mathcal{K}$ a set of *admissible* kernels, the data-driven selection of the optimal ones can be related to the following double optimization:

$$\inf_{k\in\mathcal{K}} \left\{ \inf_{f\in\mathcal{H}_k} F(f) + 2\tau \|f\|_k^2 \right\} \tag{1}$$

where $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) with kernel k, and $F : \mathcal{H}_k \to \mathbb{R}_+$ is some error functional, a very popular choice being the square loss, which we use below.

If $\mathcal{K}$ is the convex hull of a finite number of basic kernels $k_1, \ldots, k_M$, a major contribution of [7] is to show that an optimal solution of (1) can be selected in the product space of all the RKHS $\mathcal{H}_{k_j}$ corresponding to each k_j. More specifically, the authors show that the optimization problem (1) is equivalent to the following:

$$\min \left\{ F(f) + 2\tau \sum_{j=1}^{M} \|f_j\|_{k_j} : \; f = \sum_{j=1}^{M} f_j, f_j \in \mathcal{H}_{k_j} \right\}, \tag{2}$$

where $\|\cdot\|_{k_j}$ denotes the norm in $\mathcal{H}_{k_j}$. Under such assumption, given a training set $\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where the input-output pairs (x_i, y_i) belong to $\mathbb{R}^d \times \mathbb{R}$, the choice of the square loss means $F(f) = F(f_1, \ldots, f_M) = \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{M} f_j(x_i) - y_i)^2$.

Actually, from the viewpoint of our work, one of the key insights that can be drawn from equation (2) is that selecting the optimal kernel amounts to finding a sparse representation of f w.r.t. the basis

kernels. We exploited such insight and formulated an algorithm in which the two contributions of the loss function and the penalty term are disjoint. As we will see below, the effect of the sum-of-norm-based penalty in equation (2) resulted in a generalization of the soft-thresholding operator proposed in [4], which iteratively selects all the relevant kernels.

In order to ease the notation let's introduce the penalty functional $J(f) = \sum_{j=1}^{M} \|f_j\|_{\mathcal{H}_{k_j}}$. Consequently the complete error functional can be written as:

$$\mathcal{E}(f) = \operatorname*{argmin}_{f \in \mathcal{H}} F(f) + 2\tau J(f). \tag{3}$$

In the next session, which represents the key contribution of this paper, we introduce an iterative algorithm which converges to the optimal solution f^* of $\mathcal{E}$. Actually, the algorithm is indeed general, in the sense that it doesn't require F to be necessarily the square loss, and J could be slightly changed or adapted too. Indeed, the only relevant properties of F and J that we used are the following. Both $F, J : \mathcal{H} \to \mathbb{R}$ must be convex functionals. F is required to be differentiable, while J is has to be one-homogeneous, i.e. $J(\lambda f) = \lambda J(f)$, for all $f \in \mathcal{H}$ and $\lambda \in \mathbb{R}^+$. By exploiting such assumptions, we are able to derive optimality conditions for f^* and to introduce the general iterative algorithm that converges to f^*.

Due to the limited length of this note, we do not present the mathematical and technical details proving the validity of the optimality conditions and the convergence of the algorithm. We remark that these facts can be rigorously proved (indeed in a much more general setting), as we will show in a forthcoming paper. We also have some preliminary yet encouraging experimental results: a rigorous discussion of them will also appear in the same paper.

2 The Algorithm

In the first part of this section we rewrite more compactly the loss term F in (3) in order to easily compute some blocks of the algorithm we propose. Let's denote by $\mathcal{H}$ the product space $\mathcal{H}_{k_1} \times \ldots \times \mathcal{H}_{k_M}$. From the well known reproducing property [2], given $f \in \mathcal{H}$ we have $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}} := \sum_{j=1}^{M} \langle f_j, k_j(x,\cdot)\rangle_{\mathcal{H}_{k_j}}$. In fact, for $j = 1, \ldots, M$, the operator $S_x^{(j)} : \mathcal{H}_{k_j} \to \mathbb{R}^n$ defined by $f_j \mapsto (\langle f_j, k_j(x_1,\cdot)\rangle_{\mathcal{H}}, \ldots, \langle f_j, k_j(x_1,\cdot)\rangle_{\mathcal{H}})$ is linear and bounded, so that denoting by $S_x : \mathcal{H}_k \to \mathbb{R}^n$ the operator defined by $(f_1, \ldots, f_M) \mapsto \sum_{j=1}^{M} S_x^{(j)} f_j$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} f_j(x_i) - y_i\right)^2 := \|S_x f - y\|_n^2 . \tag{4}$$

Although, in general, the hypotheses space associated to a multiple kernel learning problem is infinite dimensional, the minimizer of functional (3) can be shown to have a finite representation. In fact, a straightforward generalization of the one-kernel representer theorem leads to the following solution of the underlying optimization problem:

$$f^*(\cdot) = \left(\sum_{i=1}^{n} \alpha_{1,i} k_1(x_i,\cdot), \ldots, \sum_{i=1}^{n} \alpha_{m,i} k_m(x_i,\cdot)\right).$$

Moreover, introducing the following notation

$$\begin{aligned}
&\alpha = (\alpha_1, \ldots, \alpha_M) \text{ with } \alpha_j = (\alpha_{j,1}, \ldots, \alpha_{j,n}),\\
&\mathbf{k}(x) = (\mathbf{k}_1(x), \ldots, \mathbf{k}_M(x))^T \text{ with } \mathbf{k}_j(x) = (k_j(x_1,x), \ldots, k_j(x_n,x)),\\
&\mathrm{K} = (\mathrm{K}_1, \ldots, \mathrm{K}_M) \text{ with } [\mathrm{K}_j]_{ii'} = k_j(x_i, x_i').
\end{aligned}$$

we can write the above solution as $f^*(x) = ({\alpha_1^*}^T \mathbf{k}_1(x), \ldots, {\alpha_m^*}^T \mathbf{k}_m(x))$, and we can restrict the algorithm to functions having the same expression of f^*.

Now that we are equipped with the notation, let's introduce the *MKL algorithm*, in which the operator $\hat{\mathbf{S}}_{\tau/\sigma}$ is the above mentioned generalization of soft-thresholding (see below for details). In

the second part of the section we show that the solution f^* of equation (3) can be indeed computed through such iterative algorithm.

Algorithm 1 MKL Algorithm

set $\alpha^0 = 0$
for $p = 1, 2, \ldots,$ `MAX_ITER` **do**

$$\alpha^p = \hat{\mathbf{S}}_{\tau/\sigma}\left(K, (\alpha^{p-1} - \frac{1}{\sigma n}\mathrm{K}^T(\mathrm{K}\alpha^{p-1} - y))\right)$$

end for

return $\left(\alpha^{\texttt{MAX\_ITER}}\right)^T \mathbf{k}$.

First, we provide some motivations to explain where the structure of the algorithm comes from and to state a rigorous result on its convergence. Thanks to the one-homogeneity property of J already noted in the introduction, it is easy to compute the Fenchel conjugate of J. Moreover, relying on the simple structure of the Fenchel conjugate of J, we get that for every $\sigma > 0$ the minimizer f^* of $\mathcal{E}$ satisfies the following fixed point equation:

$$f^* = (I - \pi_{\frac{\tau}{\sigma}K})(f^* - \frac{1}{2\sigma}\nabla F(f^*)), \tag{5}$$

where $K := \partial J(0)$ is the subdifferential of J at zero ([5]), and $\pi_{\frac{\tau}{\sigma}K}$ denotes the projection onto the convex set $\frac{\tau}{\sigma}K$. This suggests to solve problem (3) via the fixed point iteration

$$f^p = (I - \pi_{\frac{\tau}{\sigma}K})(f^{p-1} - \frac{1}{2\sigma}\nabla F(f^{p-1})) \tag{6}$$

for a suitable value of the parameter σ.

Each iteration can be splitted in two separate steps. The first one consists in the evaluation of the quantity $f^{p-1} - \frac{1}{2\sigma}\nabla F(f^{p-1})$ and depends only on the loss term F, while the second involves the computation of the projection onto a convex set, entirely characterized by J. Let's now show that the projection on $\frac{\tau}{\sigma}K$ can be explicitly computed.

Assuming that $f(x) = ({\alpha_1}^T\mathbf{k}_1(x), \ldots, {\alpha_m}^T\mathbf{k}_m(x))$, and recalling equation (4), we can easily compute $\nabla F(f)$ in terms of α, obtaining:

$$\nabla F(f) = \left(\left(\alpha - \frac{1}{\sigma n}\mathrm{K}^T\left(\mathrm{K}\alpha - y\right)\right)^T \mathbf{k}\right).$$

The final *ingredient* of the algorithm is the projection onto K. To this purpose, we recall that, by standard results in convex analysis, it is possible to show that the subdifferential of the penalty term can be decomposed block-wise as: $\partial J(0) = (\partial J_1(0), \ldots, \partial J_m(0))$, where $J_j : \mathcal{H}_{K_j} \to \mathbb{R}$ acts as $J_j(f_j) = \|f_j\|_{\mathcal{H}_{K_j}}$. Moreover it is easy to show that $\partial J_j(0) = \{v \in \mathcal{H}_{K_j} : \|v\|_{\mathcal{H}_{K_j}} \leq 1\}$.

We thus get that the projection is defined as $\pi_{\lambda K}(g) = \lambda\bar{v}$ with

$$\bar{v} = \operatorname*{argmin}_{v\in\mathcal{H}, \|v_j\|_{\mathcal{H}_k}\leq 1} \|\lambda v - g\|^2_{\mathcal{H}},$$

which can be computed block-wise as

$$\bar{v}_j = \min\left\{1, \frac{\|g_j\|}{\lambda}\right\}\frac{g_j}{\|g_j\|} = \min\left\{1, \frac{\sqrt{\alpha_j^T\mathrm{K}_j\alpha_j}}{\lambda}\right\}\frac{\alpha_j^T\mathbf{k}_j}{\sqrt{\alpha_j^T\mathrm{K}_j\alpha_j}}$$

when $g = (\alpha_1 \cdot \mathbf{k}_1, \ldots, \alpha_M \cdot \mathbf{k}_M)$. Finally, the operation $(I - \pi_{\lambda K})(g)$ is then given by

$$(I-\pi_{\lambda K})(g) = \left(\frac{\alpha_1^T\mathbf{k}_1}{\sqrt{\alpha_1^T\mathrm{K}_1\alpha_1}}(\sqrt{\alpha_1^T\mathrm{K}_1\alpha_1} - \lambda)_+, \ldots, \frac{\alpha_M^T\mathbf{k}_M}{\sqrt{\alpha_M^T\mathrm{K}_M\alpha_M}}(\sqrt{\alpha_M^T\mathrm{K}_M\alpha_M} - \lambda)_+\right) := \hat{\mathbf{S}}_\lambda(\mathrm{K}, \alpha)^T\mathbf{k}.$$

The convergence of the MKL algorithm can be rigorously proved even in a more general infinite dimensional setting, as the following theorem shows.

Theorem 1. *Let $F, J : \mathcal{H} \to [0, +\infty]$ be convex functionals. Assume morever that F is strictly convex and differntiable, and that J is one-homeogeneus and coercive and let f^* be the unique minimizer of $\mathcal{E} := F + J$. Then there exists $\sigma > 0$ such that algorithm 6 converges to the solution f^* for every starting point f_0.*

Remark 1. *Strict convexity of the term F may be a too strong assumption in the context of multiple kernel learning. We remark that to ensure strict convexity it is enough to add the term $\mu \|f\|^2_{\mathcal{H}}$, with $\mu > 0$ arbitrarily small.*

3 Concluding Remarks

In this section, in order to promote the discussion, we make some concluding remarks about the properties of the algorithm for learning the kernel.

- to our knowledge, in the specific case of differentiable loss functions such as the square loss, the approach we propose in the paper to solve the optimization problem 3 is the first algorithm which can be proved to converge to the optimal solution. Indeed, since we do not make use of greedy steps in the algorithm to make the optimization problem more tractable, possible approximation errors – w.r.t. the optimal solution – are due only to an early stopping of the algorithm before actual convergence is reached.
- despite most of the discussions above have been made using the square loss, the algorithm is well suited also to a number of different differentiable losses used in practice in machine learning such as, e.g., the exponential loss and the logistic loss.
- The formal requirements stated above for the penalty term – i.e. to be a convex, (possibly) non-differentiable one-homogenous functional – are met indeed by several important learning schemes. Specifically, it is easy to verify that our assumptions are met whenever the penalty can be defined through a sum of norms in distinct Euclidean spaces: $J(f) = \sum_{k=1}^{p} ||\mathcal{J}_k(f)||$, where, for all k, $\mathcal{J}_k : \mathcal{H} \to \mathbb{R}^{m_k}$ is a bounded linear operator and $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^{m_k}$. Therefore, our algorithm can be generalized straightforwardly to a more general class of problems, such as feature selection and multitask lerning.

In conclusion, we believe the proposed algorithmic framework to solve multiple kernel learning problems – as well as the mathematical background underlying it – could be of interest to the audience of the workshop. The MKL algorithm is guaranteed to converge to an optimal solution and it is easy to implement. Moreover, the same procedure can be adapted to solve also different interesting learning problems. Further developments are needed to make the algorithm usable extensively in practice. As (immediate) future works we plan to study methods to seed up the convergence of the algorithm and to make it available on the web a prototype implementation.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2006.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[3] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[4] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[5] I. Ekeland and R. Temam. *Convex analysis and variational problems*. North-Holland Publishing Co., Amsterdam, 1976.

[6] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, November 2004.

[7] Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *J. Mach. Learn. Res.*, 6:1099–1125, 2005.