社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Regularization Strategies and Empirical Bayesian Learning for MKL

Ryota TOMIOKA† and Taiji SUZUKI†

† Department of Mathematical Informatics, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.

Abstract Multiple kernel learning (MKL) has received considerable attention recently. In this paper, we show how different MKL algorithms can be understood as applications of different types of regularization on the kernel weights. We show that many algorithms based on Ivanov regularization, have their corresponding Tikhonov regularization formulations. In addition, we show that the two regularization strategies are connected by the block-norm formulation. The Tikhonov-regularization-based formulation of MKL allows us to consider a generative probabilistic model behind MKL. Based on this model, we propose learning algorithms for the kernel weights through the maximization of marginalized likelihood.

Key words Multiple Kernel Learning (MKL), Regularization, Empirical Bayesian Learning, Evidence

1. Introduction

In many learning problems, the choice of feature representation, descriptors, or kernels plays a crucial role.

The optimal representation is problem specific. For example, we can represent a web page as a bag-of-words, which might help us in classifying whether the page is discussing politics or economy; we can also represent the same page by the links provided in the page, which could be more useful in classifying whether the page is supporting a political party A or B. Similarly in a visual categorization task, a color-based descriptor might be useful in classifying an apple from a lemon but not in discriminating an airplane from a car.

Given that there is no single feature representation that works in every learning problem, it is crucial to combine them in a problem dependent manner for a successful data analysis.

In this paper, we consider the problem of combining multiple data sources in a kernel-based learning framework. The use of kernels allows us to integrate heterogeneous information (numerical features, texts, links) in an unified manner. Although feature selection has always been an important issue in statistical data analysis, our focus will be on combining features to achieve better discriminative power but not necessarily on simplifying the classifier by reducing the number of features.

More specifically, we assume that a data point $x \in \mathcal{X}$ lies in a space $\mathcal{X}$ and we are given M candidate kernel functions $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ $(m = 1, \ldots, M)$. Each kernel function corresponds to one data source. A conical combination of k_m $(m = 1, \ldots, M)$ gives the combined kernel function $\bar{k} = \sum_{m=1}^{M} d_m k_m$, where d_m is a nonnegative weight. Our goal is to find a good set of kernel weights based on some training examples.

Recall that a kernel function defines a function space called reproducing kernel Hilbert space (RKHS) [1], [2]. Thus in our setting we are given M RKHSs $\mathcal{H}_1, \ldots, \mathcal{H}_M$. For a kernel learning algorithm, the representer theorem guarantees that the best classifier in the RKHS can be represented by finite number of parameters. This is also the case when we learn M functions $f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M$ as we see in the next section.

Various approaches have been proposed for the above problem under the name multiple kernel learning (MKL) [3], [4]. The first contribution of this paper is to understand these approaches as application of different regularization strategies on the kernel weights. The pioneering work by Lanckriet et al. [3] used a simplex constraint on the kernel weights, which Kloft et al. [5] viewed as a form of Ivanov regularization and extended it to general ℓ_p-norm constraint on the kernel weights; see also Cortes et al. [6]. The Elastic-net MKL proposed in Tomioka & Suzuki [7] (see also [8], [9]) was inspired by the block-norm reformulation [10] of the work of Lanckriet et al. [3]. In this paper, we show that the Elastic-net MKL can also be understood as an application of Ivanov and Tikhonov regularizations.

The second contribution of this paper is to derive an empirical Bayesian learning algorithm for MKL motivated by the Tikhonov regularization formulation. Although Bayesian approaches have been applied to MKL earlier in a transductive nonparametric setting [11], and a setting similar to the relevance vector machine [12] in [13], [14], we believe that our formulation is more coherent with the correspondence between Gaussian process classification/regression and kernel methods [15]. In addition, we propose two iterative algorithms for the learning of kernel weights through the maximization of marginalized likelihood. One algorithm iteratively solves a reweighted MKL problem and the other iterates between a classifier training for a fixed kernel combination and a kernel weight update. Kloft et al. [16] has also recently shown that many MKL algorithms can be understood in a particular block-norm formulation that is related to Ivanov regularization. However, its connection to a probabilistic generative model is unclear.

This paper is organized as follows. In Sec. 2, we first analyze learning with fixed kernel combination and present a representer theorem. In Sec. 3, we discuss three strategies, namely Ivanov regularization, Tikhonov regularization, and generalized block norm formulation and their relations. In Sec. 4, extending the Tikhonov regularization view, we propose an empirical Bayesian approach to kernel learning. Finally, we summarize our contributions in Sec. 5.

2. Learning with fixed kernel combination

We assume that we are given N training examples $(x_i, y_i)_{i=1}^N$ where x_i belongs to an input space $\mathcal{X}$ and y_i belongs to an output space $\mathcal{Y}$ (usual settings are $\mathcal{Y} = \{\pm 1\}$ for classification and $\mathcal{Y} = \mathbb{R}$ for regression).

We first consider a learning problem with fixed kernel weights. More specifically, we fix non-negative kernel weights $d_1, d_2, \ldots, d_M$ and consider the RKHS $\bar{\mathcal{H}}$ corresponding to the combined kernel function $\bar{k} = \sum_{m=1}^M d_m k_m$. The squared RKHS norm of a function $\bar{f}$ in the combined RKHS $\bar{\mathcal{H}}$ can be represented as follows:

$$\|\bar{f}\|_{\bar{\mathcal{H}}}^2 := \min_{\substack{f_1 \in \mathcal{H}_1, \\ \ldots, f_M \in \mathcal{H}_M}} \sum_{m=1}^M \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} \quad \text{s.t. } \bar{f} = \sum_{m=1}^M f_m, \quad (1)$$

where $\mathcal{H}_m$ is the RKHS that corresponds to the kernel function k_m. See Sec 6 in [2], and also [17] for the proof. We also provide some intuition for a finite dimensional case in Appendix 1.

This important representation seems to have been overlooked in the MKL community, although it appeared in Micchelli & Pontil [17] and also in the very classical paper by Aronszajn [2]. For example, in Zien & Ong [18], the above expression was introduced as a trick to make the optimization problem convex.

Using the above representation, a supervised learning problem with a fixed kernel combination can be written as follows:

$$\underset{\substack{f_1 \in \mathcal{H}_1, \\ \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}}}{\text{minimize}} \sum_{i=1}^N \ell\left(y_i, \textstyle\sum_{m=1}^M f_m(x_i) + b\right) + \frac{C}{2} \sum_{m=1}^M \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m}, \quad (2)$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function and we assume that ℓ is convex in the second argument; for example, the loss function can be the hinge loss

$$\ell_H(y_i, z_i) = \max(0, 1 - y_i z_i),$$

or the quadratic loss

$$\ell_Q(y_i, z_i) = \frac{(y_i - z_i)^2}{\sigma_y^2}.$$

It might seem that we are making the problem unnecessarily complex by introducing M functions f_m to optimize instead of simply optimizing over $\bar{f}$. However, explicitly handling the kernel weights enables us to consider various regularization strategies on the weights as we see in the next section.

The representer theorem [1] holds for the learning problem (2), and importantly, the expansion coefficients are the same for all functions f_m (except the kernel weight d_m). In order to see this, we take the Fréchet derivative of the objective (2) and set it to zero as follows:

$$\left\langle h_m, -\sum_{i=1}^N \alpha_i k_m(\cdot, x_i) + C\frac{f_m}{d_m} \right\rangle_{\mathcal{H}_m} = 0 \qquad (\forall h_m \in \mathcal{H}_m, \forall m),$$

$$\sum_{i=1}^N \alpha_i = 0,$$

$$\partial \ell(y_i, \textstyle\sum_{m=1}^M f_m(x_i) + b) \ni -\alpha_i \qquad (i = 1, \ldots, N),$$

where $\partial \ell$ denotes the subdifferential of the loss function ℓ with respect to the second argument. From the first equation, we have the kernel expansion

$$f_m(x) = \frac{d_m}{C} \sum_{i=1}^N \alpha_i k_m(x, x_i) \qquad (m = 1, \ldots, M),$$

from which the overall predictor can be written as follows:

$$\bar{f}(x) + b = \frac{1}{C} \sum_{i=1}^N \alpha_i \sum_{m=1}^M d_m k_m(x, x_i) + b.$$

3. Learning kernel weights

Now we are ready to also optimize the kernel weights d_m in the above formulation.

MKL model	g	h	μ
block 1-norm MKL	$\sqrt{x}$	d_m	1
ℓ_p-norm MKL	$\frac{1+p}{2p}x^{p/(1+p)}$	d_m^p	$1/p$
Uniform-weight MKL (block 2-norm MKL)	$x/2$	$I_{[0,1]}(d_m)$	$+0$
block q-norm MKL ($q>2$)	$\frac{1}{q}x^{q/2}$	$d_m^{-q/(q-2)}$	$-(q-2)/q$
Elastic-net MKL	$(1-\lambda)\sqrt{x}+\frac{\lambda}{2}x$	$\frac{(1-\lambda)d_m}{1-\lambda d_m}$	$1-\lambda$

Table 1 Correspondence of the concave function g in the block-norm formulation (9), and the regularizer h and constant μ in the Ivanov and Tikhonov formulations (3) and (6). $I_{[0,1]}$ denotes the indicator function of the interval $[0,1]$; i.e., $I_{[0,1]}(x)=0$ (if $x\in[0,1]$), and $I_{[0,1]}(x)=\infty$ (otherwise).

Clearly there is a need for regularization, because the objective (2) is a monotone decreasing function of the kernel weights d_m. Intuitively speaking, d_m corresponds to the complexity allowed for the mth regression function f_m; the more complexity we allow, the better the fit to the training examples becomes. Thus without any constraint on d_m, we can get a severe overfitting problem.

There are essentially two ways to prevent such overfitting. One is to enforce some constraints on d_m, which is called Ivanov regularization and the other is to add a penalty term to the objective, which is called Tikhonov regularization; see also [5].

In the next two subsections, we show that the two regularization strategies on the kernel weights can be reduced to an equivalent block-norm formulation without kernel weights. Moreover, we extend one of the block-norm formulations and show how this can be related back to the above two regularization strategies.

Table 1 summarizes the regularization strategies we discuss in this section.

3.1 Ivanov regularization

One way to penalize the complexity is to enforce some constraint on the kernel weights for the minimization of the objective (2) as follows (see [4], [18]∼[20]):

$$\begin{aligned}\underset{\substack{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,\\ b\in\mathbb{R},\\ d_1\geqq 0,\dots,d_M\geqq 0}}{\text{minimize}} & \sum_{i=1}^{N}\ell\left(y_i,\textstyle\sum_{m=1}^{M}f_m(x_i)+b\right)\\ &+\frac{C}{2}\sum_{m=1}^{M}\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} \qquad (3)\\ \text{s.t.}\quad & \sum_{m=1}^{M}h(d_m)\leqq 1,\end{aligned}$$

where $h(d_m)$ is a convex increasing function over the non-negative reals.

The above formulation reduces to the following block 1-norm formulation in the special case when $h(d_m)=d_m$ as follows:

$$\begin{aligned}\underset{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,b\in\mathbb{R}}{\text{minimize}} & \sum_{i=1}^{N}\ell\left(y_i,\textstyle\sum_{m=1}^{M}f_m(x_i)+b\right)\\ &+\frac{C}{2}\left(\sum\nolimits_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}\right)^2, \qquad (4)\end{aligned}$$

because of Jensen's inequality; in fact,

$$\begin{aligned}\sum_{m=1}^{M}\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} &= \left(\textstyle\sum_{m=1}^{M}d_m\right)\sum_{m=1}^{M}\frac{d_m}{\sum_{m=1}^{M}d_m}\left(\frac{\|f_m\|_{\mathcal{H}_m}}{d_m}\right)^2\\ &\geqq \left(\textstyle\sum_{m=1}^{M}d_m\right)\sum_{m=1}^{M}\left(\frac{\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}}{\sum_{m=1}^{M}d_m}\right)^2\\ &\geqq \left(\sum\nolimits_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}\right)^2,\end{aligned}$$

where we used Jensen's inequality in the second line, and the equality is obtained by taking $d_m=\|f_m\|_{\mathcal{H}_m}/\sum_{m'=1}^{M}\|f_{m'}\|_{\mathcal{H}_{m'}}$.

The MKL using the simplex constraint ($h(d_m)=d_m$) or equivalently the block 1-norm MKL (often called ℓ_1-MKL) has been recently criticized because it typically results in a overly sparse solution (only few non-zero d_m) and does not necessarily perform better than simply setting $d_m=1$ for all $m=1,\dots,M$ (uniform weight combination) [21].

The ℓ_p-norm MKL proposed by Kloft et al. [5] (see also [17]) can be obtained by choosing the regularizer $h(d_m)$ as $h(d_m)=d_m^p$. In fact,

$$\begin{aligned}\sum_{m=1}^{M}\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} &= \left(\textstyle\sum_{m=1}^{M}d_m^p\right)\sum_{m=1}^{M}\frac{d_m^p}{\sum_{m=1}^{M}d_m^p}\left(\frac{\|f_m\|_{\mathcal{H}_m}^{\frac{2p}{1+p}}}{d_m^p}\right)^{\frac{1+p}{p}}\\ &\geqq \left(\textstyle\sum_{m=1}^{M}d_m^p\right)\left(\frac{\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}^{\frac{2p}{1+p}}}{\sum_{m=1}^{M}d_m^p}\right)^{\frac{1+p}{p}}\\ &\geqq \left(\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}^{\frac{2p}{1+p}}\right)^{\frac{1+p}{p}},\end{aligned}$$

where the equality is obtained by taking $d_m\propto\|f_m\|_{\mathcal{H}_m}^{2/(1+p)}$ with the normalization $(\sum_{m=1}^{M}d_m^p)^{1/p}=1$. Accordingly, we have the block-norm formulation of the ℓ_p-norm MKL as follows:

$$\begin{aligned}\underset{f_1\in\mathcal{H}_1,\dots,f_M\in\mathcal{H}_M,b\in\mathbb{R}}{\text{minimize}} & \sum_{i=1}^{N}\ell\left(y_i,\textstyle\sum_{m=1}^{M}f_m(x_i)+b\right)\\ &+\frac{C}{2}\left(\sum\nolimits_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}^q\right)^{2/q}, \qquad (5)\end{aligned}$$

where we define $q:=2p/(1+p)$. Note that the regularization term is the squared block q-norm of the regression functions f_m; see also [5], [16].

The mapping between p and q is a little bit special; $p=1$ corresponds to $q=1$, which is the block 1-norm formulation (4); $p=\infty$ corresponds to $q=2$. Note that $q=2$

corresponds to the uniform-weight combination, because the regularization term in Eq. (5) is simply the sum of squared RKHS norms, which equals Eq. (2) with $d_m = 1$ for all $m = 1, \ldots, M$. Therefore, the ℓ_p-norm MKL smoothly interpolates between the sparse block 1-norm MKL and the uniform weight combination.

Note that we can also derive the block q-norm regularization with $q > 2$ considered in Nath et al. [22] by reversing the above relation. In fact, we obtain the block q-norm formulation (5) ($q > 2$) by choosing the regularizer $h(d_m) = d_m^p$ with $p := q/(2-q)$ and using the equality constraint $\sum_{m=1}^M h(d_m) = 1$ instead of the inequality constraint in (3). However the intuition that $h(d_m)$ is a regularizer does not hold anymore, because for $q > 2$, p is negative and $h(d_m)$ is a decreasing function.

3. 2 Tikhonov regularization

Another way to penalize the complexity is to minimize the objective (2) together with the regularizer $h(d_m)$ as follows:

$$\underset{\substack{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}, \\ d_1 \geqq 0, \ldots, d_M \geqq 0}}{\text{minimize}} \sum_{i=1}^N \ell\left(y_i, \textstyle\sum_{m=1}^M f_m(x_i) + b\right) + \frac{\tilde{C}}{2} \sum_{m=1}^M \left(\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \mu h(d_m) \right), \tag{6}$$

where the regularization constant $\mu > 0$ is introduced to make a correspondence between the above formulation to the Ivanov-regularization-based formulation we discussed in the previous subsection.

A block q-norm formulation, which is equivalent to (5), can be obtained by choosing $h(d_m) = d_m^p$ and $\mu = 1/p$. In fact,

$$\frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \frac{1}{p} d_m^p = \frac{1+p}{p} \left(\frac{p}{1+p} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \frac{1}{1+p} d_m^p \right) \geqq \frac{1+p}{p} \|f_m\|_{\mathcal{H}_m}^{2p/(1+p)} = \frac{1+p}{p} \|f_m\|_{\mathcal{H}_m}^q,$$

where we used Young's inequality, which is the inequality of arithmetic and geometric means when $p = 1$; the equality is obtained by taking $d_m = \|f_m\|_{\mathcal{H}_m}^{2/(1+p)}$. The resulting block-norm formulation can be written as follows:

$$\underset{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, b \in \mathbb{R}}{\text{minimize}} \sum_{i=1}^N \ell\left(y_i, \textstyle\sum_{m=1}^M f_m(x_i) + b\right) + \frac{\tilde{C}}{q} \sum\nolimits_{m=1}^M \|f_m\|_{\mathcal{H}_m}^q, \tag{7}$$

Although the above optimization problem takes the qth power of the block q-norm of the regression functions f_m instead of the square, the two problems (5) and (7) can be mapped to each other by suitably converting C and $\tilde{C}$.

It is now clear from the above expression that block 1-norm MKL ($q = 1$) is nothing but a kernelized group lasso [23], [24].

Varma & Ray [25] considered a linear regularization on the kernel weights, which is similar to (6). For $p = 1$ ($q = 1$), the block-norm formulation (7) was considered earlier in [10].

Let us consider the block q-norm MKL for $q > 2$ of Nath et al. [22] in the Tikhonov regularization framework. Nath et al.'s approach can be interpreted as a nonconvex regularization on the kernel weights. The easiest way to see this is to extrapolate the mapping between p and q also for $q > 2$, which gives the regularization term $\mu h(d_m)$ as follows:

$$\mu h(d_m) = -\frac{q-2}{q} d_m^{-q/(q-2)}. \tag{8}$$

This is a concave increasing function. Young's inequality cannot be used to see how the above regularizer (8) is related to the block q-norm regularization, because $p = -q/(q-2)$ is negative. However, by explicitly computing the minimum, we have for $2 < q < \infty$,

$$\frac{\|f_m\|_{\mathcal{H}_m}}{d_m} - \frac{q-2}{q} d_m^{-q/(q-2)} \geqq \frac{2}{q} \|f_m\|_{\mathcal{H}_m}^q,$$

where the minimum is obtained for $d_m = \|f_m\|_{\mathcal{H}_m}^{2-q}$.

Regularization methods that do not necessarily belong to either Ivanov or Tikhonov regularization have also been proposed. Longworth & Gales [8] proposed to penalize the squared sum of kernel weights $\sum_{m=1}^M d_m^2/2$ together with a simplex constraint on the kernel weights. Their method is equivalent to the elastic-net regularization [7], [9] (see next subsection). It is worth noting that their method cannot be obtained by simply defining the regularizer $h(d_m)$ as the sum of linear and quadratic terms. This is because the simplex constraint $\sum_{m=1}^M d_m = 1$ is stronger than the inequality constraint in the Ivanov regularization problem (3) in the previous subsection.

The Tikhonov regularization formulation (6) allows a probabilistic interpretation as a hierarchical maximum a posteriori (MAP) estimation problem. The loss term can be considered as a negative log-likelihood. The first regularization term $\|f_m\|_{\mathcal{H}_m}^2/d_m$ can be considered as the negative log of a Gaussian process prior with variance scaled by the hyperparameter d_m. The last regularization term $\mu h(d_m)$ corresponds to the negative log of a hyper-prior distribution $p(d_m) \propto \exp(-\mu h(d_m))$. Instead of a MAP estimation, we can maximize the marginalized likelihood (evidence) to obtain the kernel weights. See Sec. 4. for an evidence maximization algorithm for MKL using the quadratic loss.

3. 3 Generalized block-norm formulation

The block-norm formulation (7) can be extended by using a concave function g as follows:

$$\underset{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, b \in \mathbb{R}}{\text{minimize}} \sum_{i=1}^N \ell\left(y_i, \textstyle\sum_{m=1}^M f_m(x_i) + b\right)$$

$$+ C \sum_{m=1}^{M} g(\|f_m\|^2_{\mathcal{H}_m}), \tag{9}$$

where g is a smooth concave function defined on the non-negative reals, and we assume that $\tilde{g}(x) = g(x^2)$ is a convex function of x. For example, taking $g(x) = \sqrt{x}$ gives the block 1-norm MKL (Eq. (7) with $q = 1$) and taking $g(x) = x^{q/2}/q$ gives the block q-norm MKL in Eq. (7). Note that the word "block-norm" is used with a slight abuse, because for a general concave function g, the regularizer in Eq. (9) is no longer a norm.

Why does g have to be concave? Because for a concave function g, we can relate the generalized formulation (9) back as a Tikhonov regularization problem (6) using a convex upper-bounding technique. In fact, for any given concave function g, the *concave conjugate* g^* of g is defined as follows:

$$g^*(y) = \inf_{x \geqq 0} \left(xy - g(x) \right). \tag{10}$$

The definition of concave conjugate immediately implies that the following inequality is true:

$$g\left(\|f_m\|^2_{\mathcal{H}_m}\right) \leqq \frac{\|f_m\|^2_{\mathcal{H}_m}}{2d_m} - g^*\left(\frac{1}{2d_m}\right).$$

Note that the equality is obtained by taking $d_m = 1/(2g'(\|f_m\|^2_{\mathcal{H}_m}))$, where g' is the derivative of g. Comparing the above expression to the regularization term in the Tikhonov regularization problem (6), we have

$$\mu h(d_m) = -2g^*\left(\frac{1}{2d_m}\right). \tag{11}$$

In Tomioka & Suzuki [7], the following elastic-net regularizer g was considered:

$$g(x) = (1-\lambda)\sqrt{x} + \frac{\lambda}{2}x. \tag{12}$$

With the above function g, Eq. (9) becomes

$$\underset{\substack{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}}}{\text{minimize}} \sum_{i=1}^{N} \ell\left(y_i, \textstyle\sum_{m=1}^{M} f_m(x_i) + b\right) + C \sum_{m=1}^{M} \left((1-\lambda)\|f_m\|_{\mathcal{H}_m} + \frac{\lambda}{2}\|f_m\|^2_{\mathcal{H}_m} \right), \tag{13}$$

which reduces to the block 1-norm regularization ($q = 1$ in Eq. (7)) for $\lambda = 0$ and the uniform-weight combination ($q = 2$ in Eq. (7)) for $\lambda = 1$.

In order to obtain the Tikhonov regularization problem (6) corresponding to the elastic-net regularization (13), we only need to compute the relation (11) for the concave function (12). In fact, it is easy to obtain

$$\mu h(d_m) = \frac{(1-\lambda)^2 d_m}{1 - \lambda d_m}.$$

On the other hand, to obtain the Ivanov regularization problem (3) corresponding to the elastic-net regularization (13), we need to identify the function h (without the constant μ). Choosing $h(d_m) = (1-\tilde{\lambda})d_m/(1-\tilde{\lambda}d_m)$ (note that $\tilde{\lambda}$ is different from λ), the regularization term in the Ivanov regularization problem (3) can be written as

$$\begin{aligned} \sum_{m=1}^{M} \frac{\|f_m\|^2_{\mathcal{H}}}{d_m} &= \sum_{m=1}^{M} \frac{1 - \tilde{\lambda}d_m + \tilde{\lambda}d_m}{d_m} \|f_m\|^2_{\mathcal{H}} \\ &= \sum_{m=1}^{M} \left(\frac{1-\tilde{\lambda}}{h(d_m)} + \tilde{\lambda} \right) \|f_m\|^2_{\mathcal{H}} \\ &\geqq (1-\tilde{\lambda}) \left(\textstyle\sum_{m=1}^{M} \|f_m\|_{\mathcal{H}} \right)^2 + \tilde{\lambda} \sum_{m=1}^{M} \|f_m\|^2_{\mathcal{H}}, \end{aligned}$$

where we used Jensen's inequality in the last line. The Ivanov regularization problem (3) with the above regularizer $h(d_m)$ is equivalent to the elastic-net problem (13) by suitably converting the pair (C, λ) and $(\tilde{C}, \tilde{\lambda})$.

4. Bayesian multiple kernel learning

Let us rewrite the Tikhonov regularization problem (6) as a probabilistic generative model as follows:

$$\begin{aligned} d_m &\sim \frac{1}{Z_1(\mu)} \exp(-\mu h(d_m)) \qquad (m = 1, \ldots, M), \\ f_m &\sim GP(f_m; 0, d_m k_m) \qquad (m = 1, \ldots, M) \\ b &\sim \mathcal{N}(b; 0, \sigma_b^2), \\ \bar{f} &= f_1 + \cdots + f_M, \\ y_i &\sim \frac{1}{Z_2} \exp(-\ell(y_i, \bar{f}(x_i) + b)), \end{aligned}$$

where $Z_1(\mu)$ and Z_2 are normalization constants; $GP(f; 0, k)$ denotes the Gaussian process [15] with mean zero and covariance function k; the prior variance σ_b^2 can be set very high to recover Eq. (6) in which no regularization term on the bias term exists.

When the loss function is quadratic

$$\ell(y_i, z_i) = \frac{1}{2\sigma_y^2}(y_i - z_i)^2,$$

we can analytically integrate out the Gaussian process random variable $(f_m)_{m=1}^M$ and compute the negative log of the marginalized likelihood as follows:

$$-\log p(\boldsymbol{y}|\boldsymbol{d}) = \frac{1}{2}\log\left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right| + \frac{1}{2}\boldsymbol{y}^\top \bar{\boldsymbol{K}}(\boldsymbol{d})^{-1}\boldsymbol{y} \tag{14}$$

where $\boldsymbol{d} = (d_1, \ldots, d_M)^\top$, $\boldsymbol{K}_m = (k_m(x_i, x_j))_{i,j=1}^N$ is the Gram matrix, and

$$\bar{\boldsymbol{K}}(\boldsymbol{d}) := \sigma_y^2 \boldsymbol{I}_N + \sum_{m=1}^{M} d_m \boldsymbol{K}_m.$$

We assume $\sigma_b^2 = 0$ for simplicity. In the sequel, we also omit the hyper-prior on the kernel weights d_m.

We could directly maximize (e.g., by gradient descent) the marginalized likelihood (14) with the hyper-prior term $\mu h(d_m)$ to obtain a hyperparameter maximum likelihood estimation. However this could be challenging because of the nonconvexity of the marginalized likelihood.

We present two alternative approaches for the maximization of the marginalized likelihood (14) below. The first approach is based on upper-bounding both terms in Eq. (14); since the upper-bound takes a form of the Tikhonov regularization problem (6), we can minimize this efficiently using various algorithms for MKL proposed recently [19], [26], [27]. The second approach uses the same upper-bound on the quadratic term in Eq. (14) but leaves the log determinant term as it is. Then we perform a fixed-point iteration known as the MacKay update [28], [29] for the optimization of the kernel weights.

It is easy to see that the quadratic term in the negative log-likelihood (14) can be upper bounded as follows (see [29]):

$$\frac{1}{2}\boldsymbol{y}^\top \bar{\boldsymbol{K}}(\boldsymbol{d})^{-1}\boldsymbol{y} \leqq \frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum_{m=1}^M \boldsymbol{f}_m\right\|^2 + \frac{1}{2}\sum_{m=1}^M \frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{d_m}, \tag{15}$$

where $\boldsymbol{f}_m := (f_m(x_1),\ldots,f_m(x_N))^\top$, and $\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2 = \boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1}\boldsymbol{f}_m$. Note that the above expression corresponds to the first two terms in the Tikhonov regularization problem (6).

Next, we upper-bound the log determinant term. First noticing that the function $\psi(\boldsymbol{d}) := \log\left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right|$ is concave in d_m (see p73 in [30]), we have

$$\log\left|\bar{\boldsymbol{K}}(\boldsymbol{d})\right| \leqq \sum_{m=1}^M z_m d_m - \psi^*(\boldsymbol{z}), \tag{16}$$

where $\boldsymbol{z} = (z_1,\ldots,z_M) \in \mathbb{R}^M$ and ψ^* is the concave conjugate function of ψ (see Eq. (10)). See [29], [31] for the details and other approaches (upper-bound and lower-bound) to approximate the log determinant term.

Combining the two upper-bounds (15) and (16), we have

$$-\log p(\boldsymbol{y}|\boldsymbol{d}) \leqq \frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum_{m=1}^M \boldsymbol{f}_m\right\|^2 + \frac{1}{2}\sum_{m=1}^M\left(\frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{d_m} + z_m d_m\right) - \frac{1}{2}\psi^*(\boldsymbol{z}).$$

Comparing the above expression to the Tikhonov problem (6), we can see that minimization of the right-hand side with respect to $(\boldsymbol{f}_m)_{m=1}^M$ and $\boldsymbol{d}$ is a weighted block 1-norm MKL. In fact, by explicitly minimizing over $\boldsymbol{d}$, we have

$$\min_{\boldsymbol{d}} -\log p(\boldsymbol{y}|\boldsymbol{d}) \leqq \frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum_{m=1}^M \boldsymbol{f}_m\right\|^2 + \sum_{m=1}^M \sqrt{z_m}\,\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m} - \frac{1}{2}\psi^*(\boldsymbol{z}).$$

Once we solve the block 1-norm MKL for a fixed variational parameter $\boldsymbol{z}$, we can minimize over $\boldsymbol{z}$ to tighten the upper-bound. Accordingly the iteration can be written as follows:

$$(\boldsymbol{f}_m)_{m=1}^M \leftarrow \operatorname*{argmin}_{(\boldsymbol{f}_m)_{m=1}^M}\left(\frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum\nolimits_{m=1}^M \boldsymbol{f}_m\right\|^2 + \sum\nolimits_{m=1}^M \sqrt{z_m}\,\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}\right),$$
$$z_m \leftarrow \operatorname{Tr}\left((\sigma_y^2\boldsymbol{I}_N + \textstyle\sum_{m=1}^M d_m\boldsymbol{K}_m)^{-1}\boldsymbol{K}_m\right).$$

The second approach computes the derivative of the negative log likelihood to derive a fixed-point iteration. By minimizing the left-hand side of Eq. (15), we have

$$-\log p(\boldsymbol{y}|\boldsymbol{d}) = \frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum_{m=1}^M \boldsymbol{f}_m^{\mathrm{FKL}}\right\|^2 + \frac{1}{2}\sum_{m=1}^M \frac{\|\boldsymbol{f}_m^{\mathrm{FKL}}\|_{\boldsymbol{K}_m}^2}{d_m} + \frac{1}{2}\log\left|\sigma_y^2\boldsymbol{I}_N + \sum_{m=1}^M d_m\boldsymbol{K}_m\right|,$$

where $\boldsymbol{f}_m^{\mathrm{FKL}}$ is the minimizer of the upper-bound (15); note that this minimization is a fixed kernel weight learning problem (2).

Taking the derivative of the above expression with respect to d_m we have

$$-\frac{\|\boldsymbol{f}_m^{\mathrm{FKL}}\|_{\boldsymbol{K}_m}^2}{d_m^2} + \operatorname{Tr}\left((\sigma^2\boldsymbol{I}_N + \textstyle\sum_{m=1}^M d_m\boldsymbol{K}_m)^{-1}\boldsymbol{K}_m\right) = 0.$$

Therefore, we use the following iteration:

$$(\boldsymbol{f}_m)_{m=1}^M \leftarrow \operatorname*{argmin}_{(f_m)_{m=1}^M}\left(\frac{1}{2\sigma_y^2}\left\|\boldsymbol{y}-\sum\nolimits_{m=1}^M \boldsymbol{f}_m\right\|^2 + \frac{1}{2}\sum\nolimits_{m=1}^M \frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{d_m}\right)$$
$$d_m \leftarrow \frac{\|\boldsymbol{f}_m\|_{\boldsymbol{K}_m}^2}{\operatorname{Tr}\left((\sigma^2\boldsymbol{I}_N + \sum_{m=1}^M d_m\boldsymbol{K}_m)^{-1}d_m\boldsymbol{K}_m\right)}.$$

Note that the update equation for the kernel weight d_m is neither unique nor the most simple one, although the denominator can be interpreted as the effective number of parameters [28]. The convergence of this procedure is not established mathematically, but it is known to converge rapidly in many practical situations [12]. Wipf & Nagarajan [29] proposed an alternative update rule that can be shown to converge.

5. Summary

We have shown that various MKL algorithms including ℓ_p-norm MKL and Elastic-net MKL can be seen as applications of different regularization strategies. We have shown that the three formulations, Ivanov regularization, Tikhonov

regularization, and the generalized block-norm formulation can be transformed to each other; see Table 1. The Tikhonov regularization-based formulation allows us to view MKL as a hierarchical Gaussian process model. Motivated by this view, we have shown that the marginalized likelihood can be maximized by simple iterative algorithms that iteratively solves a reweighted block 1-norm MKL or a fixed kernel weight learning problem. Further analysis and empirical validation are necessary to gain more insights about the proposed empirical Bayesian learning procedure.

Acknowledgement We would like to thank Hisashi Kashima and Shinichi Nakajima for helpful discussions. This work was partially supported by MEXT Kakenhi 22700138, 22700289.

Appendix

1. Proof of Eq. (1) in a finite dimensional case

In this section, we provide a proof of Eq. (1) when $\mathcal{H}_1, \ldots, \mathcal{H}_m$ are all finite dimensional. We assume that the input space $\mathcal{X}$ consists of N points $x_1, \ldots, x_N$, for example the training points. The function $f_m \in \mathcal{H}_m$ is completely specified by the function values at the N-points $\boldsymbol{f}_m = (f_m(x_1), \ldots, f_m(x_N))^\top$. The kernel function k_m is also specified by the Gram matrix $\boldsymbol{K}_m = (k_m(x_i, x_j))_{i,j=1}^N$. The inner product $\langle f_m, g_m \rangle_{\mathcal{H}_m}$ is written as $\langle f_m, g_m \rangle_{\mathcal{H}_m} = \boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1} \boldsymbol{g}_m$, where $\boldsymbol{g}_m$ is the N-dimensional vector representation of $g_m \in \mathcal{H}_m$, assuming that the Gram matrix $\boldsymbol{K}_m$ is positive definite. It is easy to check the reproducibility; in fact, $\langle f_m, k_m(\cdot, x_i) \rangle = \boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1} \boldsymbol{K}_m(:, i) = f(x_i)$, where $\boldsymbol{K}_m(:, i)$ is a column vector of the Gram matrix $\boldsymbol{K}_m$ that corresponds to the ith sample point x_i.

The right-hand side of Eq. (1) is written as follows:

$$\min_{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M \in \mathbb{R}^N} \sum_{m=1}^M \frac{\boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1} \boldsymbol{f}_m}{d_m} \quad \text{s.t.} \quad \sum_{m=1}^M \boldsymbol{f}_m = \bar{\boldsymbol{f}}.$$

Forming the Lagrangian, we have

$$\begin{aligned}
&\sum_{m=1}^M \frac{\boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1} \boldsymbol{f}_m}{d_m} \\
&= \sum_{m=1}^M \frac{\boldsymbol{f}_m^\top \boldsymbol{K}_m^{-1} \boldsymbol{f}_m}{d_m} + 2\boldsymbol{\alpha}^\top \Big(\bar{\boldsymbol{f}} - \sum\nolimits_{m=1}^M \boldsymbol{f}_m \Big) \\
&\geqq -\boldsymbol{\alpha}^\top \left(\sum\nolimits_{m=1}^M d_m \boldsymbol{K}_m \right) \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \bar{\boldsymbol{f}} \\
&\xrightarrow{\max_{\boldsymbol{\alpha}}} \bar{\boldsymbol{f}}^\top \left(\sum\nolimits_{m=1}^M d_m \boldsymbol{K}_m \right)^{-1} \bar{\boldsymbol{f}},
\end{aligned}$$

where the equality is obtained for

$$\boldsymbol{f}_m = d_m \boldsymbol{K}_m \left(\sum\nolimits_{m=1}^M d_m \boldsymbol{K}_m \right)^{-1} \bar{\boldsymbol{f}}.$$

References

[1] B. Schölkopf and A. Smola: "Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond", MIT Press, Cambridge, MA (2002).

[2] N. Aronszajn: "Theory of reproducing kernels", Transactions of the American Mathematical Society, **68**, pp. 337–404 (1950).

[3] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett and M. Jordan: "Learning the kernel matrix with semi-definite programming", Journal of Machine Learning Research, **5**, pp. 27–72 (2004).

[4] F. Bach, G. Lanckriet and M. Jordan: "Multiple kernel learning, conic duality, and the SMO algorithm", the 21st International Conference on Machine Learning, pp. 41–48 (2004).

[5] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller and A. Zien: "Efficient and accurate lp-norm multiple kernel learning", Advances in Neural Information Processing Systems 22 (Eds. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta), pp. 997–1005 (2009).

[6] C. Cortes, M. Mohri and A. Rostamizadeh: "L_2 regularization for learning kernels", the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009) (2009). Montréal, Canada.

[7] R. Tomioka and T. Suzuki: "Sparsity-accuracy trade-off in MKL", Technical report, arXiv:1001.2615 (2010).

[8] C. Longworth and M. Gales: "Combining derivative and parametric kernels for speaker verification", IEEE Transactions on Audio, Speech, and Language Processing, **17**, 4, pp. 748–757 (2009).

[9] J. Shawe-Taylor: "Kernel learning for novelty detection", In NIPS 08 Workshop: Kernel Learning – Automatic Selection of Optimal Kernels (2008).

[10] F. R. Bach, R. Thibaux and M. I. Jordan: "Computing regularization paths for learning multiple kernels", Advances in Neural Information Processing Systems 17, MIT Press, pp. 73–80 (2005).

[11] Z. Zhang, D. Yeung and J. Kwok: "Bayesian inference for transductive learning of kernel matrix using the tanner-wong data augmentation algorithm", Proceedings of the Twenty-First International Conference on Machine LearningACM, p. 118 (2004).

[12] M. E. Tipping: "Sparse bayesian learning and the relevance vector machine", J. Mach. Learn. Res., **1**, pp. 211–244 (2001).

[13] M. Girolami and S. Rogers: "Hierarchic bayesian models for kernel learning", Proceedings of the 22nd International Conference on Machine LearningACM, pp. 241–248 (2005).

[14] T. Damoulas and M. A. Girolami: "Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection", Bioinformatics, **24**, 10, pp. 1264–1270 (2008).

[15] C. E. Rasmussen and C. K. I. Williams: "Gaussian Processes for Machine Learning", MIT Press (2006).

[16] M. Kloft, U. Rückert and P. L. Bartlett: "A unifying view of multiple kernel learning", Technical report, arXiv:1005.0437 (2010).

[17] C. A. Micchelli and M. Pontil: "Learning the kernel function via regularization", Journal of Machine Learning Research, **6**, pp. 1099–1125 (2005).

[18] A. Zien and C. Ong: "Multiclass multiple kernel learning", Proceedings of the 24th international conference on machine learningACM, pp. 11910–1198 (2007).

[19] S. Sonnenburg, G. Rätsch, C. Schäfer and B. Schölkopf: "Large scale multiple kernel learning", Journal of Machine Learning Research, **7**, pp. 1531–1565 (2006).

[20] A. Rakotomamonjy, F. Bach, S. Canu and G. Y.: "Simplemkl", Journal of Machine Learning Research, **9**, pp. 2491–2521 (2008).

[21] C. Cortes: "Can learning kernels help performance?", In-

vited talk at International Conference on Machine Learning (ICML 2009). Montréal, Canada (2009).

[22] J. S. Nath, G. Dinesh, S. Raman, C. Bhattacharyya, A. Ben-Tal and K. R. Ramakrishnan: "On the algorithmics and applications of a mixed-norm based kernel learning formulation", Advances in Neural Information Processing Systems, **22**, pp. 844–852 (2009).

[23] M. Yuan and Y. Lin: "Model selection and estimation in regression with grouped variables", Journal of The Royal Statistical Society Series B, **68**, 1, pp. 49–67 (2006).

[24] F. R. Bach: "Consistency of the group lasso and multiple kernel learning", Journal of Machine Learning Research, **9**, pp. 1179–1225 (2008).

[25] M. Varma and D. Ray: "Learning the discriminative power-invariance trade-off", IEEE 11th International Conference on Computer Vision (ICCV), pp. 1–8 (2007).

[26] O. Chapelle and A. Rakotomamonjy: "Second order optimization of kernel parameters", NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels, Whistler (2008).

[27] T. Suzuki and R. Tomioka: "SpicyMKL", Technical report, arXiv:0909.5026 (2009).

[28] D. J. C. MacKay: "Bayesian interpolation", Neural computation, **4**, 3, pp. 415–447 (1992).

[29] D. Wipf and S. Nagarajan: "A unified bayesian framework for meg/eeg source imaging", NeuroImage, **44**, 3, pp. 947–966 (2009).

[30] S. Boyd and L. Vandenberghe: "Convex Optimization", Cambridge University Press, Cambridge (2004).

[31] M. Seeger and H. Nickisch: "Large scale variational inference and experimental design for sparse generalized linear models", Technical report, arXiv:0810.0901 (2008).