Regularization Strategies and Empirical Bayesian Learning for MKL

Ryota Tomioka\(^1\), Taiji Suzuki\(^1\)

\(^1\)Department of Mathematical Informatics, The University of Tokyo

2010-12-11
NIPS2010 Workshop:
New Directions in Multiple Kernel Learning
Our contribution

- Relationships between **different regularization strategies**
  - Ivanov regularization (kernel weights)
  - Tikhonov regularization (kernel weights)
  - (Generalized) block-norm formulation (no kernel weights)

  Are they equivalent? — in which way?

- Empirical Bayesian learning algorithm for MKL
  - Maximizes the marginalized likelihood
  - Can be considered as a **non-separable regularization** on the kernel weights.
Learning with a fixed kernel combination

Fixed kernel combination $k_d(x, x') = \sum_{m=1}^{M} d_m k_m(x, x')$.

$$\minimize_{\tilde{f} \in \mathcal{H}(d), \quad b \in \mathbb{R}} \sum_{i=1}^{N} \ell \left( y_i, \tilde{f}(x_i) + b \right) + \frac{C}{2} \| \tilde{f} \|_{\mathcal{H}(d)}^2,$$

($\mathcal{H}(d)$ is the RKHS corresponding to the combined kernel $k_d$) is equivalent to learning $M$ functions $(f_1, \ldots, f_M)$ as follows:

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

where $\tilde{f}(x) = \sum_{m=1}^{M} f_m(x)$.

See Sec. 6 in Aronszajn (1950), Micchelli & Pontil (2005).
Regularization Strategies

Ivanov regularization

We can *constrain* the size of kernel weights $d_m$ by

$$\minimize_{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \atop b \in \mathbb{R}, \atop d_1 \geq 0, \ldots, d_M \geq 0} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{2} \sum_{m=1}^{M} \left\| f_m \right\|_{\mathcal{H}_m}^2,$$

$$\text{s.t.} \quad \sum_{m=1}^{M} h(d_m) \leq 1 \quad (h \text{ is convex, increasing}).$$

Equivalent to the more common expression:

$$\minimize_{f \in \mathcal{H}(d), \atop b \in \mathbb{R}, \atop d_1 \geq 0, \ldots, d_M \geq 0} \sum_{i=1}^{N} \ell \left( y_i, f(x_i) + b \right) + \frac{C}{2} \left\| f \right\|_{\mathcal{H}(d)}^2, \quad \text{s.t.} \quad \sum_{m=1}^{M} h(d_m) \leq 1.$$
Tikhonov regularization

We can *penalize* the size of kernel weights $d_m$ by

$$\minimize_{\begin{array}{l} f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \\ b \in \mathbb{R}, \\ d_1 \geq 0, \ldots, d_M \geq 0 \end{array}} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right)$$

$$+ \frac{C}{2} \sum_{m=1}^{M} \left( \frac{\| f_m \|^2_{\mathcal{H}_m}}{d_m} + \mu h(d_m) \right).$$

Note that the above is equivalent to

$$\minimize_{\begin{array}{l} f \in \mathcal{H}(d), \\ b \in \mathbb{R}, \\ d_1 \geq 0, \ldots, d_M \geq 0 \end{array}} \sum_{i=1}^{N} \ell \left( y_i, f(x_i) + b \right) + \frac{C}{2} \| f \|^2_{\mathcal{H}(d)} + \frac{C\mu}{2} \sum_{m=1}^{M} h(d_m).$$

\[\text{data-fit} \quad \text{f-prior} \quad \text{d}_m\text{-hyper-prior}\]
Are these two formulations equivalent?

Previously thought that...
Yes. But the choice of the pair \((C, \mu)\) is complicated.
\[
\Rightarrow \text{In the Tikhonov formulation we have to choose both } C \text{ and } \mu!
\]
(Kloft et al., 2010)

We show that...
If you give up the constant 1 in the Ivanov formulation
\[
\sum_{m=1}^{M} h(d_m) \leq 1,
\]
- Correspondence via equivalent block-norm formulations.
- \(C\) and \(\mu\) can be chosen independently.
- The constant 1 has no meaning.
Ivanov ⇒ block-norm formulation 1 (known)

Let \( h(d_m) = d_m^p \) (\( \ell_p \)-norm MKL); see Kloft et al. (2010).

\[
\begin{aligned}
\text{minimize} & \quad \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{2} \sum_{m=1}^{M} \frac{\|f_m\|^2_{\mathcal{H}_m}}{d_m} , \\
\text{s.t.} & \quad \sum_{m=1}^{M} d_m^p \leq 1 .
\end{aligned}
\]

\[\downarrow\quad \text{Jensen’s inequality}\]

\[
\begin{aligned}
\text{minimize} & \quad \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{2} \left( \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^q \right)^{2/q} ,
\end{aligned}
\]

where \( q = \frac{2p}{1 + p} \). Minimum is attained at \( d_m \propto \|f_m\|_{\mathcal{H}_m}^{2/(1+p)} \).
**Tikhonov ⇒ block-norm formulation 2 (new)**

Let \( h(d_m) = d_m^p \), \( \mu = 1/p \) (\( \ell_p \)-norm MKL)

\[
\text{minimize } \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{2} \sum_{m=1}^{M} \left( \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + \frac{d_m^p}{p} \right).
\]

\[\downarrow\] Young's inequality

\[
\text{minimize } \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{q} \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}^q.
\]

where \( q = 2p/(1 + p) \). Minimum is attained at \( d_m = \|f_m\|_{\mathcal{H}_m}^{2/(1+p)} \).
The two block norm formulations are equivalent

Block norm formulation 1 (from Ivanov):

$$\min_{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, \bin \in \mathbb{R}} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{\tilde{C}}{2} \left( \sum_{m=1}^{M} \| f_m \|_{\mathcal{H}_m}^q \right)^{2/q}.$$ 

Block norm formulation 2 (from Tikhonov):

$$\min_{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M, b \in \mathbb{R}} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + \frac{C}{q} \sum_{m=1}^{M} \| f_m \|_{\mathcal{H}_m}^q.$$ 

- Just have to map $C$ and $\tilde{C}$.
- The implied kernel weights are normalized/unnormalized.
Generalized block-norm formulation

\[
\text{minimize} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) + C \sum_{m=1}^{M} g(\|f_m\|_{\mathcal{H}_m}^2), \quad (4)
\]

where \( g \) is a concave block-norm-based regularizer.

Example (Elastic-net MKL): \( g(x) = (1 - \lambda)\sqrt{x} + \frac{\lambda}{2} x, \)

\[
\text{minimize} \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) + b \right) \]

\[
+ C \sum_{m=1}^{M} \left( (1 - \lambda)\|f_m\|_{\mathcal{H}_m} + \frac{\lambda}{2} \|f_m\|_{\mathcal{H}_m}^2 \right),
\]

where \( \mathcal{H}_m \) is a set of functions.
Theorem

Correspondence between the convex (kernel-weight-based) regularizer $h(d_m)$ and the concave (block-norm-based) regularizer $g(x)$ is given as follows:

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

where $g^*$ is the concave conjugate of $g$.

Proof: Use the concavity of $g$ as

$$\frac{\|f_m\|_{\mathcal{H}_m}^2}{2d_m} \geq g(\|f_m\|_{\mathcal{H}_m}^2) + g^*(1/(2d_m)).$$

See also Palmer et al. (2006).
Examples

Generalized Young's inequality:

\[ xy \geq g(x) + g^*(y) \]

where \( g \) is concave, and \( g^* \) is the concave conjugate of \( g \).

Example 1: let \( g(x) = \sqrt{x} \), then \( g^*(y) = -1/(4y) \) and

\[
\frac{\|f_m\|_{H_m}^2}{2d_m} + \frac{d_m}{2} \geq \|f_m\|_{H_m} \quad \text{(L1-MKL)}.
\]

Example 2: let \( g(x) = x^{q/2}/q \) (\( 1 \leq q \leq 2 \)), then \( g^*(y) = \frac{q-2}{2q} (2y)^{q/(q-2)} \)

\[
\frac{\|f_m\|_{H_m}^2}{2d_m} + \frac{d_m^p}{2p} \geq \frac{1}{q} \|f_m\|_{H_m}^q \quad \text{(\( \ell_p \)-norm MKL)},
\]

where \( p := q/(2 - q) \).
## Correspondence

<table>
<thead>
<tr>
<th>MKL model</th>
<th>block-norm ( g(x) )</th>
<th>kern weight ( h(d_m) )</th>
<th>reg const ( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>block 1-norm MKL</td>
<td>( \sqrt{x} )</td>
<td>( d_m )</td>
<td>1</td>
</tr>
<tr>
<td>( \ell_p )-norm MKL</td>
<td>( \frac{1+p}{2p} x^{p/(1+p)} )</td>
<td>( d_m^p )</td>
<td>( 1/p )</td>
</tr>
<tr>
<td>Uniform-weight MKL (block 2-norm MKL)</td>
<td>( x/2 )</td>
<td>( I_{[0,1]}(d_m) )</td>
<td>( +0 )</td>
</tr>
<tr>
<td>block ( q )-norm MKL (( q &gt; 2 ))</td>
<td>( \frac{1}{q} x^{q/2} )</td>
<td>( d_m^{-q/(q-2)} )</td>
<td>( -(q-2)/q )</td>
</tr>
<tr>
<td>Elastic-net MKL</td>
<td>((1 - \lambda)\sqrt{x} + \frac{\lambda}{2} x)</td>
<td>( \frac{(1-\lambda)d_m}{1-\lambda d_m} )</td>
<td>( 1 - \lambda )</td>
</tr>
</tbody>
</table>

\( I_{[0,1]}(x) \) is the indicator function of the closed interval \([0, 1]\); i.e., \( I_{[0,1]}(x) = 0 \) if \( x \in [0, 1] \), and \( +\infty \) otherwise.
Empirical Bayesian MKL

Bayesian view

Tikhonov regularization as a hierarchical MAP estimation

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) \right) + \sum_{m=1}^{M} \frac{\| f_m \|_{\mathcal{H}_m}^2}{2d_m} + \mu \sum_{m=1}^{M} h(d_m). \\
\text{likelihood} & \quad \text{f}_m\text{-prior} \quad \text{d}_m\text{-hyper-prior}
\end{align*}
\]

Hyper prior over the kernel weights

\[
d_m \sim \frac{1}{Z_1(\mu)} \exp(-\mu h(d_m)) \quad (m = 1, \ldots, M).
\]

Gaussian process for the functions

\[
f_m \sim \mathcal{GP}(f_m; 0, d_m k_m) \quad (m = 1, \ldots, M).
\]

Likelihood

\[
y_i \sim \frac{1}{Z_2(x_i)} \exp(-\ell(y_i, \sum_{m=1}^{M} f_m(x_i))).
\]
Marginalized likelihood

Assume Gaussian likelihood

\[ \ell(y, z) = \frac{1}{2\sigma^2_y} (y - z)^2. \]

The marginalized likelihood (omitting hyper-prior for simplicity)

\[- \log p(y|d) \]

\[= \underbrace{\frac{1}{2\sigma^2_y} \left\| y - \sum_{m=1}^{M} f^\text{MAP}_m \right\|^2}_{\text{likelihood}} + \frac{1}{2} \sum_{m=1}^{M} \frac{\|f^\text{MAP}_m\|_{\mathcal{H}_m}^2}{d_m} + \frac{1}{2} \log |\tilde{K}(d)|. \]

- \( f^\text{MAP}_m \): MAP estimate for a fixed kernel weights \( d_m \) (\( m = 1, \ldots, M \)).
- \( \tilde{K}(d) := \sigma^2_y I_N + \sum_{m=1}^{M} d_m K_m. \)

See also Wipf & Nagarajan (2009).
Comparing MAP and empirical Bayes objectives

Hyper-prior MAP (MKL):

\[
\sum_{i=1}^{N} \ell \left( y_i, \sum_{m=1}^{M} f_m(x_i) \right) + \frac{1}{2} \sum_{m=1}^{M} \frac{\| f_m \|_H^2}{d_m} + \mu \sum_{m=1}^{M} h(d_m) .
\]

Empirical Bayes:

\[
\frac{1}{2\sigma_y^2} \left\| y - \sum_{m=1}^{M} f_{\text{MAP}}^m \right\|_2^2 + \frac{1}{2} \sum_{m=1}^{M} \frac{\| f_{\text{MAP}}^m \|_H^2}{d_m} + \frac{1}{2} \log |\tilde{K}(d)| .
\]
Caltech 101 dataset (classification)

- Regularization constant $C$ chosen by $2 \times 4$-fold cross validation on the training-set.

![Graph showing accuracy vs number of samples per class for different methods: MKL (logit), Uniform, MKL (square), ElasticnetMKL ($\lambda=0.5$), and BayesMKL.](image)
1,760 kernel functions.

- 4 SIFT features (hsvsift, sift, sift4px, sift8px)
- 22 spacial decompositions (including spatial pyramid kernel)
- 2 kernel functions (Gaussian and $\chi^2$)
- 10 kernel parameters

**Graph:**

- BayesMKL $\text{acc}=0.82$
- ElasticnetMKL ($\lambda=0.5$) $\text{acc}=0.97$
- MKL (square) $\text{acc}=0.80$
- Uniform $\text{acc}=0.92$
- MKL (logit) $\text{acc}=0.82$

**Accuracy values:**

\[ [0.82, 0.92, 0.80, 0.97, 0.82] \]
Caltech 101 dataset: kernel weights (detail)

[8.166667e−01 9.166667e−01 8.000000e−01 9.666667e−01 8.166667e−01]

MKL (logit)

Uniform

MKL (square)

chi²–kernel

Gaussian kernel

chi²

ElasticnetMKL

BayesMKL

Ryota Tomioka (Univ Tokyo)
Summary

- Two regularized kernel weight learning formulations
  - Ivanov regularization.
  - Tikhonov regularization.
  are equivalent. No additional tuning parameter!
- Both formulations reduce to block-norm formulations via Jensen’s inequality / (generalized) Young’s inequality.
- Probabilistic view of MKL: hierarchical Gaussian process model.
- Elastic-net MKL performs similarly to uniform weight MKL, but shows grouping of mutually depended kernels.
- Empirical-Bayes MKL and L1-MKL seem to make the solution overly sparse, but often they choose slightly different set of kernels.
We would like to thank Hisashi Kashima and Shinichi Nakajima for helpful discussions. This work was supported in part by MEXT KAKENHI 22700138, 22700289, and NTT Communication Science Laboratories.
A brief proof

- Minimize the Lagrangian:

\[
\min_{f_1 \in \mathcal{H}_1, \ldots, f_M \in \mathcal{H}_M} \frac{1}{2} \sum_{m=1}^M \frac{\|f_m\|^2}{\mathcal{H}_m} + \left\langle g, \bar{f} - \sum_{m=1}^M f_m \right\rangle_{\mathcal{H}(d)},
\]

where \( g \in \mathcal{H}(d) \) is a Lagrangian multiplier.

- Fréchet derivative

\[
\left\langle h_m, \frac{f_m}{d_m} - \left\langle g, k_m \right\rangle_{\mathcal{H}(d)} \right\rangle_{\mathcal{H}_m} = 0 \Rightarrow f_m(x) = \left\langle g, d_m k_m(\cdot, x) \right\rangle_{\mathcal{H}(d)}.
\]

- Maximize the dual

\[
\max_{g \in \mathcal{H}(d)} -\frac{1}{2} \|g\|^2_{\mathcal{H}(d)} + \left\langle g, \bar{f} \right\rangle_{\mathcal{H}(d)} = \frac{1}{2} \|\bar{f}\|^2_{\mathcal{H}(d)}
\]
Gehler & Nowozin. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. CVPR, 2009.
Method A: upper-bounding the log det term

- Use the upper bound
  \[
  \log |\tilde{K}(d)| \leq \sum_{m=1}^{M} z_{m} d_{m} - \psi^*(\mathbf{z})
  \]

- Eliminate the kernels weights by explicit minimization (AGM ineq.)

Update \( f_m \) as

\[
(f_m)_{m=1}^{M} \leftarrow \arg\min_{(f_m)_{m=1}^{M}} \left( \frac{1}{2\sigma^2_y} \left\| \mathbf{y} - \sum_{m=1}^{M} f_m \right\|^2 + \sum_{m=1}^{M} \sqrt{z_m} \left\| f_m \right\|_{K_m} \right)
\]

Update \( z_m \) as (tightly the upper bound)

\[
z_m \leftarrow \text{Tr} \left( \sigma^2_y \mathbf{I}_N + \sum_{m=1}^{M} d_m \mathbf{K}_m \right)^{-1} \mathbf{K}_m \),
\]

where \( d_m = \left\| f_m \right\|_{\mathcal{H}_m} / \sqrt{z_m} \).

- Each update step is a *rewighted L1-MKL problem*.
- Each update step minimizes an upper bound of the...
Method B: MacKay update

- Use the fixed point condition for the update of the weights:
\[- \frac{\|f_m^{FKL}\|_2^2}{d_m^2} + \text{Tr} \left( (\sigma^2 I_N + \sum_{m=1}^M d_m K_m)^{-1} K_m \right) = 0.\]

Update $f_m$ as
\[
(f_m)_{m=1}^M \leftarrow \arg\min_{(f_m)_{m=1}^M} \left( \frac{1}{2\sigma^2_y} \left\| y - \sum_{m=1}^M f_m \right\|^2 + \frac{1}{2} \sum_{m=1}^M \frac{\|f_m\|_2^2}{d_m} \right) \]

Update the kernel weights $d_m$ as
\[
d_m \leftarrow \frac{\|f_m\|_2^2}{\text{Tr} \left( (\sigma^2 I_N + \sum_{m=1}^M d_m K_m)^{-1} d_m K_m \right)}.\]

- Each update step is a fixed kernel weight learning problem (easy).
- Convergence empirically OK (e.g., RVM).