Structured Regularization and MKL

Guillaume Obozinski
Willow group - INRIA - ENS - Paris

New directions in MKL, NIPS workshops 2010, Whistler, December 11th 2010
What is Multiple Kernel Learning?

- A formulation that allows to learn the metric/inner product in a supervised problem?
What is Multiple Kernel Learning?

- A formulation that allows to learn the metric/inner product in a supervised problem?
- A functional space / kernelized version of sparse methods?
What is Multiple Kernel Learning?

- A formulation that allows to learn the metric/inner product in a supervised problem?
- A functional space / kernelized version of sparse methods?
- A framework for *data fusion*?
What is Multiple Kernel Learning?

- A formulation that allows to learn the metric/inner product in a supervised problem?
- A functional space / kernelized version of sparse methods?
- A framework for *data fusion*?
- A way to introduce structure in the functional space?
Learning the kernel or MKL?

Standard learning problem in a RKHS

\[
\min_{w \in \mathcal{H}} L(\Phi w) + \lambda \|w\|_\mathcal{H}
\]

Dual

\[
F(K) = \max_{\alpha} -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top K \alpha
\]

- Supervised learning problems are **convex** in the kernel matrix
Learning the kernel or MKL?

Standard learning problem in a RKHS

\[
\min_{w \in \mathcal{H}} L(\Phi w) + \lambda \|w\|_{\mathcal{H}}
\]

Dual

\[
F(K) = \max_{\alpha} -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top K \alpha
\]

- Supervised learning problems are **convex** in the kernel matrix
  
  \rightarrow \text{Learn the kernel: } \min_{K \in \mathcal{K}} F(K)

- Linear combination \rightarrow SDP (Lanckriet et al., 2004b)

\[
\min_{\eta \in \mathbb{R}^p} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \sum_i \eta_i K_i \succeq 0
\]
Learning the kernel or MKL?

Standard learning problem in a RKHS

\[ \min_{w \in \mathcal{H}} L(\Phi w) + \lambda \|w\|_{\mathcal{H}} \]

Dual

\[ F(K) = \max_{\alpha} -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top K\alpha \]

- Supervised learning problems are \textbf{convex} in the kernel matrix

\[ \rightarrow \text{ Learn the kernel: } \min_{K \in \mathcal{K}} F(K) \]

- Linear combination \(\rightarrow\) SDP (Lanckriet et al., 2004b)

\[ \min_{\eta \in \mathbb{R}^p} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \sum_i \eta_i K_i \succeq 0 \]

- Convex combination \(\rightarrow\) QCQP (Lanckriet et al., 2004a)

\[ \min_{\eta \in \mathbb{R}_+^p} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \sum_i \eta_i = 1 \]
A primal for MKL (Bach et al., 2004)

Let \( w = (w_1, \ldots, w_p) \in \mathbb{R}^d \)

\[
\min_{w \in \mathbb{R}^d} L(Xw) + \frac{\lambda}{2} \left( \sum_j \|w_j\|_2 \right)^2
\]

\[
\min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}^p} L(Xw) + \frac{\lambda}{2} \sum_j \frac{\|w_j\|_2^2}{\eta_j}
\]

\[
\min_{\tilde{w} \in \mathbb{R}^d, \eta \in \mathbb{R}^p} L\left( \sum_j \eta_j^{1/2} X_j \tilde{w}_j \right) + \frac{\lambda}{2} \|\tilde{w}_j\|_2^2
\]

\[
\min_{\eta \in \mathbb{R}^p} \max_{\alpha \in \mathbb{R}^n} -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top \left( \sum_j \eta_j K_j \right)
\]

- MKL is directly related through duality with \( \ell_1 \) and \( \ell_1/\ell_2 \).
- MKL should be expected to behave like sparse methods.
**Functional interpretation:** Generalized additive models.

\[ y = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) + \epsilon \]

(Lin and Zhang, 2006; Ravikumar et al., 2008)
**Functional interpretation**: Generalized additive models.

\[ y = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) + \epsilon \]

(Lin and Zhang, 2006; Ravikumar et al., 2008)

**Data Fusion** (Lanckriet et al., 2004a)

- Provides an appropriate embedding of heterogeneous data types in the same functional space
- One of the initial selling point of MKL
MKL and Kernel Selection

- Goal? Data Fusion? Aggregation or Selection?
- Claimed that “Solution is sparse $\rightarrow$ discards irrelevant information”.
- Issue of whether MKL works in practice / improves over unweighted linear combination of kernels.
MKL and Kernel Selection

- Goal? Data Fusion? Aggregation or Selection?
- Claimed that “Solution is sparse → discards irrelevant information”.
- Issue of whether MKL works in practice / improves over unweighted linear combination of kernels.

1. Either exploit the sparsity in initial formulation
   - Initial formulation of MKL is intimately connected with sparsity.
   - is relevant to select a few kernels among a large number.

2. Or use non-sparse formulations
   - Possible to consider variants of MKL for the $\ell_p$-norms $p > 1$
     (Aflalo et al., 2010)(Kloft et al., 2009).
Structured Sparsity

Usual sparsity:

→ **cardinality** of the support: number of selected variables/non-zero parameters

Yuan and Lin (2007), Zhao et al. (2009), Baraniuk et al. (2008), Bach (2008), Jacob et al. (2009a), Jenatton et al. (2009), Jenatton et al. (2010c), He and Carin (2009), Huang et al. (2009), Jenatton et al. (2010b), Mairal et al. (2010).

→ constrains the structure of the sparsity pattern.

Examples:

- Variables should be selected in groups
- Variables lie in a hierarchy and selected respecting a partial order
- Variables lie on a graph and connected variables are likely to be simultaneously relevant.
Structured Sparsity

**Usual sparsity:**
→ **cardinality** of the support: number of selected variables/non-zero parameters

**Structured sparsity:**
Yuan and Lin (2007), Zhao et al. (2009), Baraniuk et al. (2008), Bach (2008), Jacob et al. (2009a), Jenatton et al. (2009), Jenatton et al. (2010c), He and Carin (2009), Huang et al. (2009), Jenatton et al. (2010b), Mairal et al. (2010).

→ **constrains the structure** of the sparsity pattern.
Structured Sparsity

Usual sparsity:
→ *cardinality* of the support: number of selected variables/non-zero parameters

Structured sparsity:
Yuan and Lin (2007), Zhao et al. (2009), Baraniuk et al. (2008), Bach (2008), Jacob et al. (2009a), Jenatton et al. (2009), Jenatton et al. (2010c), He and Carin (2009), Huang et al. (2009), Jenatton et al. (2010b), Mairal et al. (2010).
→ *constrains the structure* of the sparsity pattern.

Examples:
- Variables should be *selected in groups*
- Variables lie in a hierarchy and selected *respecting a partial order*.
- Variables lie on a graph and *connected variables* are likely to be simultaneously relevant.
Group Lasso extensions

Group Lasso
Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ be a partition of $\{1, \ldots, p\}$ into disjoint groups

$$\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|_q$$

- Sets to 0 groups of variables
- Support is a union of groups of variables
Group Lasso extensions

Group Lasso
Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ be a partition of $\{1, \ldots, p\}$ into disjoint groups

$$\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|_q$$

- Sets to 0 groups of variables
- Support is a union of groups of variables

Group Lasso with overlapping groups (Jenatton et al., 2009)
- $\Omega$ is still a norm
- Set of zeros is a union of groups $\bigcup_{g \in \mathcal{G}_0} g$.
- Allowed patterns: intersections of complements $\bigcap_{g \in \mathcal{G}_0} g^c$
- Can construct $\Omega$ for families of supports stable by intersection
**Group Lasso extensions**  (Jacob et al., 2009b)

**Latent Group Lasso**

- Idea: Introduce a latent variable $v_g$ per group s.t.
  - $\text{Supp}(v_g) \subset g$
  - $w = \sum_{g \in G} v_g$

$$\Omega(w) = \min_{v_1, \ldots, v_K} \sum_g \|v_g\|_q \text{ s.t. } \sum_g v_g = w$$
**Group Lasso extensions**  (Jacob et al., 2009b)

**Latent Group Lasso**

- Idea: Introduce a latent variable $v_g$ per group s.t.
  - $\text{Supp}(v_g) \subset g$
  - $w = \sum_{g \in G} v_g$

$$\Omega(w) = \min_{v_1, \ldots, v_K} \sum_{g} \|v_g\|_q \quad \text{s.t.} \quad \sum_{g} v_g = w$$

**Graph Lasso**

$$\Omega(w) = \min_{v_e} \sum_{e \in E} \|v_e\|^2 \quad \text{s.t.} \quad \sum_{g} v_g = w$$
Hierarchical Norms (Zhao et al., 2009)

Given a directed graph $(V, E)$
- $D(i)$ the set of descendants of node $i$

$$\Omega(w) = \sum_{i \in V} \|w_{D(i)}\|_2$$
Hierarchical Norms (Zhao et al., 2009)

Given a directed graph \((V, E)\)
- \(D(i)\) the set of descendants of node \(i\)

\[
\Omega(w) = \sum_{i \in V} \|w_{D(i)}\|_2
\]

Tree-structured groups
For all groups \(g\) and \(h\) in \(G\) we have

\(g \subset h\) or \(h \subset g\) or \(h \cap g = \emptyset\)

- Simple algorithms
  (Jenatton et al., 2010a)
Towards MKL...

Are there MKL counterparts for these norms?
Towards MKL...

- Are there MKL counterparts for these norms?

**Variational formulation of norms**  
(Micchelli and Pontil, 2006)

\[ \ell_1 \]
\[ \|w\|_1 \leq 1 \leq \sum_i \frac{w_i^2}{\eta_i} \]

\[ \ell_p \text{ for } 1 \leq p \leq 2 \]
\[ \|w\|_p^2 = \min_{\eta} \sum_i \frac{w_i^2}{\eta_i} \]
\[ \text{s.t. } \sum_i \eta_i \frac{p}{2-p} \leq 1 \]
Towards MKL...

- Are there MKL counterparts for these norms?

**Variational formulation of norms**
(Micchelli and Pontil, 2006)

\[ \ell_1 \]
\[ \| w \|_1^2 = \min_{\eta: \eta^\top 1 \leq 1} \sum_i \frac{w_i^2}{\eta_i} \]

\[ \ell_p \text{ for } 1 \leq p \leq 2 \]
\[ \| w \|_p^2 = \min_{\eta} \sum_i \frac{w_i^2}{\eta_i}, \text{s.t.: } \sum_i \eta_i^{\frac{p}{2-p}} \leq 1 \]

Of the form \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) for \( H \) a convex set.
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function
- \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) such that \( H \) convex set.
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function
- \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) such that \( H \) convex set.
- then \( \Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2 \).
MKL and Fenchel duality

- $L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)})$ a loss function
- $\Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j}$ such that $H$ convex set.
- then $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

$$\min_{w \in \mathbb{R}^p} L(Xw) + \frac{\lambda}{2} \Omega(w)^2$$
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function
- \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) such that \( H \) convex set.
- then \( \Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2 \).

\[
\begin{align*}
\min_{w \in \mathbb{R}^p} & \quad L(Xw) + \frac{\lambda}{2} \Omega(w)^2 \\
\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} & \quad \max_{\alpha \in \mathbb{R}^n} L(u) + \frac{\lambda}{2} \Omega(w)^2 - \alpha^T(u - Xw)
\end{align*}
\]
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function.
- \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) such that \( H \) convex set.
- then \( \Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2 \).

\[
\begin{align*}
\min_{w \in \mathbb{R}^p} & \quad L(Xw) + \frac{\lambda}{2} \Omega(w)^2 \\
\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} & \quad L(u) + \frac{\lambda}{2} \Omega(w)^2 - \alpha^T (u - Xw) \\
\max_{\alpha} & \quad \left[ \min_u L(u) - \alpha^T u \right] + \lambda \left[ \min_w \frac{1}{2} \Omega(w) + \frac{\alpha^T X}{\lambda} w \right]
\end{align*}
\]
MKL and Fenchel duality

- $L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)})$ a loss function
- $\Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j}$ such that $H$ convex set.
- then $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

$$\begin{align*}
\min_{w \in \mathbb{R}^p} & \quad L(Xw) + \frac{\lambda}{2} \Omega(w)^2 \\
\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} & \quad \max_{\alpha \in \mathbb{R}^n} L(u) + \frac{\lambda}{2} \Omega(w)^2 - \alpha^T (u - Xw) \\
\max_{\alpha} & \quad \left[ \min_u L(u) - \alpha^T u \right] + \lambda \left[ \min_w \frac{1}{2} \Omega(w) + \frac{\alpha^T X}{\lambda} w \right] \\
\max_{\alpha} & \quad -L^*(\alpha) - \frac{1}{2\lambda} \Omega^*(X^T \alpha)^2
\end{align*}$$
MKL and Fenchel duality

- \( L(Xw) = \sum_{i=1}^{n} \ell_i(w, x^{(i)}) \) a loss function
- \( \Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j} \) such that \( H \) convex set.
- then \( \Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2 \).

\[
\begin{align*}
\min_{w \in \mathbb{R}^p} & \quad L(Xw) + \frac{\lambda}{2} \Omega(w)^2 \\
\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} & \quad \max_{\alpha \in \mathbb{R}^n} \left[ \min_u L(u) - \alpha^\top u \right] + \lambda \left[ \min_w \frac{1}{2} \Omega(w) + \frac{\alpha^\top X}{\lambda} w \right] \\
\max_{\alpha} & \quad \left[ \min_u L(u) - \alpha^\top u \right] + \lambda \left[ \min_w \frac{1}{2} \Omega(w) + \frac{\alpha^\top X}{\lambda} w \right] \\
\max_{\alpha} & \quad -L^*(\alpha) - \frac{1}{2\lambda} \Omega^*(X^\top \alpha)^2
\end{align*}
\]

with \( K_j = x_j x_j^\top \) a rank one kernel.
MKL and Fenchel duality: RKHS version

Let

\[ \mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p \]
MKL and Fenchel duality: RKHS version

Let

- $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
- $w = (w_1, \ldots, w_p) \in \mathcal{B}$
MKL and Fenchel duality: RKHS version

Let

- \( \mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p \)
- \( w = (w_1, \ldots, w_p) \in \mathcal{B} \quad \phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B} \)
MKL and Fenchel duality: RKHS version

Let

- \( \mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p \)
- \( w = (w_1, \ldots, w_p) \in \mathcal{B} \quad \phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B} \)
- \( L(\Phi w) = \sum_{i=1}^{n} \ell_i(\sum_j \langle w_j, \phi_j(x(i)) \rangle_{\mathcal{H}_j}) \) a loss function
MKL and Fenchel duality: RKHS version

Let

- $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
- $w = (w_1, \ldots, w_p) \in \mathcal{B}$, $\phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B}$
- $L(\Phi w) = \sum_{i=1}^n \ell_i \left( \sum_j \langle w_j, \phi_j(x^{(i)}) \rangle_{\mathcal{H}_j} \right)$ a loss function
- $\Omega(v)^2 = \min_{\eta \in \mathcal{H}} \sum_i \frac{v_i^2}{\eta_i}$
MKL and Fenchel duality: RKHS version

Let

- $B = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
- $w = (w_1, \ldots, w_p) \in B$  \hspace{1em} $\phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in B$
- $L(\Phi w) = \sum_{i=1}^n \ell_i (\sum_j \langle w_j, \phi_j(x^{(i)}) \rangle_{\mathcal{H}_j})$  \hspace{1em} a loss function
- $\Omega(v)^2 = \min_{\eta \in \mathcal{H}} \sum_i \frac{v_i^2}{\eta_i}$

$$\min_{w \in B} L(\Phi w) + \frac{1}{2} \Omega(\|w_1\|_{\mathcal{H}_1}, \ldots, \|w_p\|_{\mathcal{H}_p})^2$$
MKL and Fenchel duality: RKHS version

Let

1. $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
2. $w = (w_1, \ldots, w_p) \in \mathcal{B}$ \quad $\phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B}$
3. $L(\Phi w) = \sum_{i=1}^n \ell_i \left( \sum_j \langle w_j, \phi_j(x^{(i)}) \rangle_{\mathcal{H}_j} \right)$ \quad a loss function
4. $\Omega(v)^2 = \min_{\eta \in \mathcal{H}} \sum_i \frac{v_i^2}{\eta_i}$

\[
\min_{w \in \mathcal{B}} \quad L(\Phi w) + \frac{1}{2} \Omega \left( \|w_1\|_{\mathcal{H}_1}, \ldots, \|w_p\|_{\mathcal{H}_p} \right)^2
\]

\[
\ldots
\]

\[
\max \min_{\alpha, \eta \in \mathcal{H}} \quad -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top \left( \sum_j \eta_j K_j \right) \alpha
\]

with $K_j = \Phi_j \Phi_j^\top$ is the kernel associated with $\mathcal{H}_j$. 
Norms and variational formulations ("\(\eta\)-trick")

\[
\| w \|_2 \quad \min_{\eta} \sum_i \frac{w_i^2}{\eta_i} \quad \text{s.t.} \quad \eta_i = 1
\]

\[
\| w \|_1 \quad \min_{\eta} \sum_i \frac{w_i^2}{\eta_i} \quad \text{s.t.} \quad \sum_i \eta_i = 1
\]

\[
\| w \|_p, \ 1 \leq p \leq 2 \quad \min_{\eta} \sum_i \frac{w_i^2}{\eta_i} \quad \text{s.t.} \quad \sum_i \eta_i^{\frac{p}{2-p}} = 1
\]

\[
\sum_{g \in G} \| w_g \|_2 \quad \min_{\eta} \sum_i \| w_g \|^2 \quad \text{s.t.} \quad \sum_i \eta_g = 1
\]

\[
\min_{v | w = \sum_g v_g} \sum_{g \in G} \| v_g \|_2 \quad \min_{\eta} \sum_i \frac{w_i^2}{\sum_g \delta_{g \ni i} \eta_g} \quad \text{s.t.} \quad \sum_g \eta_g = 1
\]

Tree latent \(l_1/l_2\)

\[
\min_{\eta} \sum_i \frac{w_i^2}{\eta_i} \quad \text{s.t.} \quad \forall j \rightarrow i, \eta_i \leq \eta_j \leq 1
\]
Multiple kernel learning schemes

\[ \| w \|_2 \quad F(K) \quad \text{with} \quad K = K_1 + \ldots + K_p \]

\[ \| w \|_1 \quad \max_{\eta} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \sum_i \eta_i = 1 \]

\[ \| w \|_p, 1 \leq p \leq 2 \quad \max_{\eta} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \sum_i \eta_i^{2-p} = 1 \]

\[ \sum_{g \in G} \| w_g \|_2 \quad \max_{\eta} F(\sum_i \frac{1}{\sum_{g \ni i} \eta_g} K_i) \quad \text{s.t.} \quad \sum_i \eta_g = 1 \]

\[ \min_{w=\sum_g v_g} \sum_{g \in G} \| v_g \|_2 \quad \max_{\eta} F(\sum_g K_g \eta_g) \quad \text{s.t.} \quad \sum_g \eta_g = 1 \]

Tree latent \( \ell_1 / \ell_2 \)

\[ \max_{\eta} F(\sum_i \eta_i K_i) \quad \text{s.t.} \quad \forall j \rightarrow i, \; \eta_i \leq \eta_j \leq 1 \]
Hierarchical kernel learning (Bach, 2008)

Decompose kernels as a large sum of “atomic” kernels indexed by a certain set $V$:

$$k(x, x') = \sum_{j \in V} k_j(x, x')$$

Nonlinear Variable Selection:
Example with $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$

- Gaussian/ANOVA kernels: $p = \#(V) = 2^q$

$$\prod_{j=1}^{q} \left( 1 + e^{-\alpha (x_j - x'_j)^2} \right) = \sum_{J \subseteq \{1, \ldots, q\}} \prod_{j \in J} e^{-\alpha (x_j - x'_j)^2} = \sum_{J \subseteq \{1, \ldots, q\}} e^{-\alpha \|x_{J} - x'_{J}\|^2_2}$$

- NB: decomposition is related to Cosso (Lin and Zhang, 2006)
Graph-based structured regularization

- $D(j)$ is the set of descendants of $j \in V$:

$$
\sum_{j \in V} d_j \| w_{D(j)} \|_2 = \sum_{j \in V} d_j \left( \sum_{i \in D(j)} \| w_i \|_2^2 \right)^{1/2}
$$

Main property If $j$ is selected, so are all its ancestors
Algorithms?

- Which algorithms can we use for structured MKL formulations?
- MKL historically challenging from an optimization point of view. Why?

\[ L(w) + \lambda \Omega(w) \]

1. Both \( L \) and \( \Omega \) smooth proximal methods, quasi-newton
2. \( \Omega \) non-smooth but “simple” proximal methods, CD
3. \( L \) non-smooth (e.g. SVM) SMO (Vishwanathan et al., 2010)
4. \( L \) and \( \Omega \) non-smooth harder!
(Sonnenburg et al., 2006; Rakotomamonjy et al., 2008; Xu et al., 2009)
Kernelized Proximal methods (Rosasco et al., 2009)

Proximal methods (Moreau, 1962), (Nesterov, 2007) (Beck and Teboulle, 2009)

Computing the proximal operator in feature space
Denote $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$ and let $w = (w_1, \ldots, w_p)$ and $u = (u_1, \ldots, u_p) \in \mathcal{B}$.

- **Proximal problem**

  $$\min_{w \in \mathcal{B}} \frac{1}{2} \| w - u \|_{\mathcal{B}} + \lambda \Omega (\| f_1 \|_{\mathcal{H}_1}, \ldots, \| f_p \|_{\mathcal{H}_p})$$

- Or, using the **representer theorem**, with $u_j = \Phi_j^T \alpha_0$, $w_j = \Phi_j^T \alpha$, $K_j = \Phi_j \Phi_j^T$ and $K = \sum_j K_j$,

  $$\min_{\alpha \in \mathbb{R}^n} (\alpha - \alpha_0)^T K (\alpha - \alpha_0) + \lambda \Omega (\alpha^T K_1 \alpha, \ldots, \alpha^T K_p \alpha)$$

- The appropriate proximity term to use:

  $\rightarrow$ the RKHS norm $\| \cdot \|_{\mathcal{B}} \leftrightarrow \alpha^T K \alpha \neq \| \alpha \|_2^2$. 
Structured sparse MKL

Sparsity is useful in the high-dimensional setting: \(\log(p) = O(n)\)
Interesting for situation with a very large number of kernel:

- Suggests to consider:
  - Combinatorial feature spaces and function spaces.
  - Hierarchical functions spaces

- Requires
  - Efficient schemes to (re)-compute the kernels on the fly
  - Kernel caching strategies
Crack the kernel!

Return to **kernel computation** algorithms and **kernel design** (Shawe-Taylor and Cristianini, 2004).

- dynamic programs to compute efficiently kernels that are sums and product of more elementary kernels.
Crack the kernel!

Return to **kernel computation** algorithms and **kernel design** (Shawe-Taylor and Cristianini, 2004).

- dynamic programs to compute efficiently kernels that are sums and product of more elementary kernels.

**All subset kernel**

For $K_J(x, y) = \prod_{j \in J} K_j(x, y)$, $K = \sum_{J \in 2^P} K_J(x, y) = \prod_{j=1}^{p} (1 + K_j(x, y))$

**Polynomial kernel** $K(x, y) = (1 + \gamma K_0(x, y))^p = \sum_{k=0}^{p} \binom{p}{k} \gamma^k K_0(x, y)^k$

**String kernel, graph kernels, pyramid match kernels**
Crack the kernel!

Return to **kernel computation** algorithms and **kernel design** (Shawe-Taylor and Cristianini, 2004).

- dynamic programs to compute efficiently kernels that are sums and product of more elementary kernels.

**All subset kernel**

For $K_J(x, y) = \prod_{j \in J} K_j(x, y)$, \[ K = \sum_{J \in 2^P} K_J(x, y) = \prod_{j=1}^{p} (1 + K_j(x, y)) \]

**Polynomial kernel** $K(x, y) = (1 + \gamma K_0(x, y))^p = \sum_{k=0}^{p} \binom{p}{k} \gamma^k K_0(x, y)^k$

**String kernel, graph kernels, pyramid match kernels**

→ Integrate MKL inside of the kernel
Conclusion

• **What is MKL?**
  - A formulation to learn in composite/structured RKHSs.
  - An opportunity to encode a priori structure of the function space.
  - Linear (conic) metric learning

• **Structured sparsity directly applicable to MKL**

• **Algorithms for structured sparsity can be applied to MKL**

• **Design algorithms to explore structured feature spaces**
References I


References II


References III


