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

$\rightarrow$ *Learn* the kernel: $\min_{K \in \mathcal{K}} F(K)$

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

$$\min_{\eta \in \mathbb{R}^p} F(\textstyle\sum_i \eta_i K_i) \quad \text{s.t.} \quad \textstyle\sum_i \eta_i K_i \succ 0$$

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 **convex** in the kernel matrix

→ *Learn* the kernel: $\min_{K \in \mathcal{K}} F(K)$

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

$$\min_{\eta \in \mathbb{R}^p} F(\textstyle\sum_i \eta_i K_i) \quad \text{s.t.} \quad \textstyle\sum_i \eta_i K_i \succ 0$$

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

$$\min_{\eta \in \mathbb{R}_+^p} F(\textstyle\sum_i \eta_i K_i) \quad \text{s.t.} \quad \textstyle\sum_i \eta_i = 1$$

A primal for MKL (Bach et al., 2004)

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

$$\min_{w \in \mathbb{R}^d} L(Xw) + \frac{\lambda}{2}\big(\textstyle\sum_j \|w_j\|_2\big)^2$$

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

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

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

- MKL is directly related through duality with ℓ_1 and ℓ_1/ℓ_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 $\rightarrow$ 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
   $\rightarrow$ Initial formulation of MKL is intimately connected with sparsity.
   $\rightarrow$ 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 ℓ_p-norms $p > 1$ (Aflalo et al., 2010)(Kloft et al., 2009).

Structured Sparsity

Usual sparsity:

$\rightarrow$ **cardinality** of the support: number of selected variables/non-zero parameters

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:

$\rightarrow$ **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).

$\rightarrow$ **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)

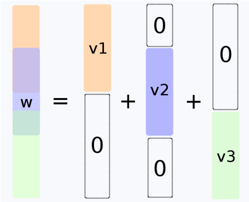
- Ω is still a norm
- Set of zeros is a union of groups $\boxed{\cup_{g \in \mathcal{G}_0} g}$.
- Allowed patterns: **intersections of complements** $\boxed{\cap_{g \in \mathcal{G}_0} g^c}$
- Can construct Ω for families of supports **stable by intersection**

Group Lasso extensions (Jacob et al., 2009b)

Latent Group Lasso

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

$$\boxed{\Omega(w) = \min_{v_1,\dots,v_K} \sum_g \|v_g\|_q \quad \text{s.t.} \quad \sum_g v_g = w}$$

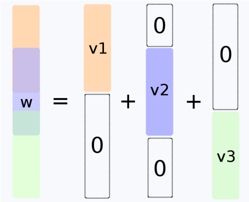


Group Lasso extensions (Jacob et al., 2009b)

Latent Group Lasso

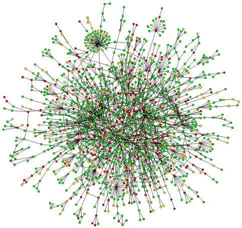
- Idea: Introduce a latent variable v_g per group s.t.
  - $\mathrm{Supp}(v_g) \subset g$
  - $w = \sum_{g \in \mathcal{G}} v_g$

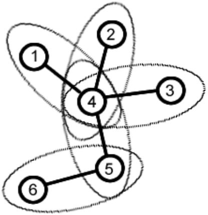
$$\boxed{\Omega(w) = \min_{v_1,\ldots,v_K} \sum_g \|v_g\|_q \quad \text{s.t.} \quad \sum_g v_g = w}$$



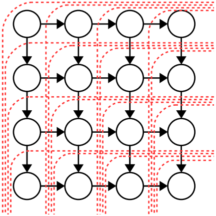
Graph Lasso

$$\Omega(w) = \min_{v_e} \quad \sum_{e \in E} \|v_e\|^2$$
$$\text{s.t.} \quad \sum_g v_g = w$$





Hierarchical Norms (Zhao et al., 2009)

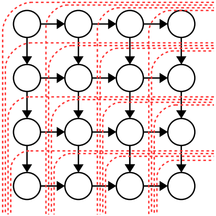


Given a directed graph (V, E)

- $D(i)$ the set of descendants of node i

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

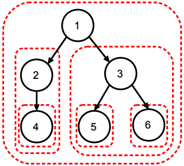
Hierarchical Norms (Zhao et al., 2009)



Given a directed graph (V, E)

- $D(i)$ the set of descendants of node i

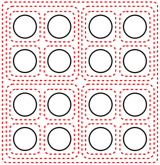
$$\Omega(w) = \sum_{i \in V} \|w_{D(i)}\|_2$$



Tree-structured groups

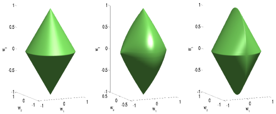
For all groups g and h in $\mathcal{G}$ we have

$$g \subset h \text{ or } h \subset g \text{ or } h \cap g = \emptyset$$



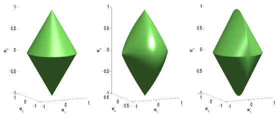
- Simple algorithms (Jenatton et al., 2010a)

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)

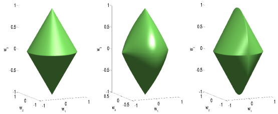
ℓ_1

$$\|w\|_1^2 = \min_{\eta:\, \eta^\top \mathbf{1} \le 1} \sum_i \frac{w_i^2}{\eta_i}$$

ℓ_p for $1 \le p \le 2$

$$\|w\|_p^2 = \min_{\eta} \sum_i \frac{w_i^2}{\eta_i}, \text{s.t. : } \sum_i \eta_i^{\frac{p}{2-p}} \le 1$$

Towards MKL...



- Are there MKL counterparts for these norms?

Variational formulation of norms
(Micchelli and Pontil, 2006)

ℓ_1

$$\|w\|_1^2 = \min_{\eta:\,\eta^\top \mathbf{1} \leq 1} \sum_i \frac{w_i^2}{\eta_i}$$

ℓ_p for $1 \leq p \leq 2$

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

Of the form $\boxed{\Omega(w)^2 = \min_{\eta \in H} \sum_j \frac{w_j^2}{\eta_j}}$ 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 $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

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

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 $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

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

$$\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \quad L(u) + \frac{\lambda}{2}\Omega(w)^2 - \alpha^\top (u - Xw)$$

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 $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

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

$$\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \quad L(u) + \frac{\lambda}{2}\Omega(w)^2 - \alpha^\top (u - Xw)$$

$$\max_{\alpha} \quad \left[\min_u L(u) - \alpha^\top u \right] + \lambda \left[\min_w \frac{1}{2}\Omega(w) + \frac{\alpha^\top X}{\lambda} w \right]$$

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 $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

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

$$\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \quad L(u) + \frac{\lambda}{2}\Omega(w)^2 - \alpha^\top (u - Xw)$$

$$\max_{\alpha} \quad \left[\min_u L(u) - \alpha^\top u\right] + \lambda \left[\min_w \frac{1}{2}\Omega(w) + \frac{\alpha^\top X}{\lambda} w\right]$$

$$\max_{\alpha} \quad -L^*(\alpha) - \frac{1}{2\lambda}\Omega^*(X^\top \alpha)^2$$

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 $\Omega^*(\kappa)^2 = \max_{\eta \in H} \sum_j \eta_j^2 \kappa_j^2$.

$$\begin{aligned}
\min_{w \in \mathbb{R}^p} \quad & L(Xw) + \frac{\lambda}{2}\Omega(w)^2 \\
\min_{w \in \mathbb{R}^p, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \quad & L(u) + \frac{\lambda}{2}\Omega(w)^2 - \alpha^\top (u - Xw) \\
\max_{\alpha} \quad & \Big[\min_u L(u) - \alpha^\top u\Big] + \lambda \Big[\min_w \frac{1}{2}\Omega(w) + \frac{\alpha^\top X}{\lambda} w\Big] \\
\max_{\alpha} \quad & -L^*(\alpha) - \frac{1}{2\lambda}\Omega^*(X^\top \alpha)^2 \\
\max_{\alpha} \min_{\eta \in H} \quad & -L^*(\alpha) - \frac{1}{2\lambda}\alpha^\top \Big(\sum_j \eta_j K_j\Big)\alpha
\end{aligned}$$

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

- $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
- $w = (w_1, \ldots, w_p) \in \mathcal{B} \qquad \phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B}$

MKL and Fenchel duality: RKHS version

Let

- $\mathcal{B} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_p$
- $w = (w_1, \ldots, w_p) \in \mathcal{B} \qquad \phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathcal{B}$
- $L(\Phi w) = \sum_{i=1}^n \ell_i(\sum_j \langle \mathbf{w}_j, \phi_j(x^{(i)}) \rangle_{\mathcal{H}_j})$ a loss function

MKL and Fenchel duality: RKHS version

Let

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

MKL and Fenchel duality: RKHS version

Let

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

$$\min_{w \in \mathcal{B}} \quad L(\Phi w) + \frac{\lambda}{2} \Omega\big(\|w_1\|_{\mathcal{H}_1}, \ldots, \|w_p\|_{\mathcal{H}_p}\big)^2$$

MKL and Fenchel duality: RKHS version

Let

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

$$\min_{w \in \mathcal{B}} \quad L(\Phi w) + \frac{\lambda}{2} \Omega\big(\|w_1\|_{\mathcal{H}_1}, \ldots, \|w_p\|_{\mathcal{H}_p}\big)^2$$

$$\ldots$$

$$\ldots$$

$$\max_{\alpha} \min_{\eta \in H} \qquad -L^*(\alpha) - \frac{1}{2\lambda} \alpha^\top \big(\sum_j \eta_j K_j\big) \alpha$$

with $K_j = \Phi_j \Phi_j^\top$ is the kernel associated with $\mathcal{H}_j$.

Norms and variational formulations ("η-trick")

$\|w\|_2$	$\min_{\eta} \sum_i \frac{w_i^2}{\eta_i}$	s.t.	$\eta_i = 1$
$\|w\|_1$	$\min_{\eta} \sum_i \frac{w_i^2}{\eta_i}$	s.t.	$\sum_i \eta_i = 1$
$\|w\|_p, \; {}_{1 \leq p \leq 2}$	$\min_{\eta} \sum_i \frac{w_i^2}{\eta_i}$	s.t.	$\sum_i \eta_i^{\frac{p}{2-p}} = 1$
$\sum_{g \in \mathcal{G}} \|w_g\|_2$	$\min_{\eta} \sum_i \frac{\|w_g\|^2}{\eta_g}$	s.t.	$\sum_i \eta_g = 1$
$\min_{\mathbf{v} \mid w = \sum_g v_g} \sum_{g \in \mathcal{G}} \|v_g\|_2$	$\min_{\eta} \sum_i \frac{w_i^2}{\sum_{g \ni i} \eta_g}$	s.t.	$\sum_g \eta_g = 1$
Tree latent ℓ_1/ℓ_2	$\min_{\eta} \sum_i \frac{w_i^2}{\eta_i}$	s.t.	$\forall j \to i, \; \eta_i \leq \eta_j \leq 1$

Multiple kernel learning schemes

$\|w\|_2$	$F(K)$	with	$K = K_1 + \ldots + K_p$
$\|w\|_1$	$\max_\eta F(\sum_i \eta_i K_i)$	s.t.	$\sum_i \eta_i = 1$
$\|w\|_{p,\ 1 \leq p \leq 2}$	$\max_\eta F(\sum_i \eta_i K_i)$	s.t.	$\sum_i \eta_i^{\frac{p}{2-p}} = 1$
$\sum_{g \in \mathcal{G}} \|w_g\|_2$	$\max_\eta F\Big(\sum_i \frac{1}{\sum_{g \ni i} \eta_g^{-1}} K_i\Big)$	s.t.	$\sum_i \eta_g = 1$
$\min_{w = \sum_g v_g} \sum_{g \in \mathcal{G}} \|v_g\|_2$	$\max_\eta F\big(\sum_g K_g \eta_g\big)$	s.t.	$\sum_g \eta_g = 1$
Tree latent ℓ_1/ℓ_2	$\max_\eta F(\sum_i \eta_i K_i)$	s.t.	$\forall j \to i,\ \eta_i \leq \eta_j \leq 1$

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 $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$

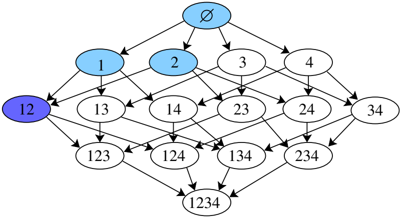
- Gaussian/ANOVA kernels: $p = \#(V) = 2^q$

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

- NB: decomposition is related to Cosso (Lin and Zhang, 2006)

Hierarchial kernel learning II (Bach, 2008)

Natural Hierarchy of Subsets: Hasse diagram



Graph-based structured regularization

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

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

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 Ω smooth	proximal methods, quasi-newton
(2)	Ω non-smooth but "simple"	proximal methods, CD
(3)	L non-smooth (e.g. SVM)	SMO (Vishwanathan et al., 2010)
(4)	L and Ω 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**

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

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

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

- The appropriate proximity term to use:
  $\rightarrow$ the RKHS norm $\|\cdot\|_{\mathcal{B}} \leftrightarrow \alpha^\top K \alpha \quad \neq \quad \|\alpha\|_2^2$.

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

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

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

$\rightarrow$ **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

Aflalo, J., Ben-Tal, A., Bhattacharyya, C., Nath, J., and Raman, S. (2010). Variable Sparsity Kernel Learning. *JMLR*. Submitted.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3):243–272.

Bach, F. (2008). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*.

Bach, F., Lanckriet, G., and Jordan, M. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. (2008). Model-based compressive sensing. Technical report, arXiv:0808.3572.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

He, L. and Carin, L. (2009). Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497.

Huang, J., Zhang, T., and Metaxas, D. (2009). Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009a). Group lasso with overlaps and graph lasso. In Bottou, L. and Littman, M., editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 433–440, Montreal. Omnipress.

References II

Jacob, L., Obozinski, G., and Vert, J.-P. (2009b). Group Lasso with overlaps and graph Lasso. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Jenatton, R., Audibert, J., and Bach, F. (2009). Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2010a). Proximal methods for hierarchical sparse coding. Technical report, arXiv:1009.3139. submitted.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2010b). Proximal methods for sparse hierarchical dictionary learning. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, Haifa, Israel. Omnipress.

Jenatton, R., Obozinski, G., and Bach, F. (2010c). Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., M
"uller, K., and Zien, A. (2009). Efficient and accurate lp-norm multiple kernel learning. *Advances in neural information processing systems*, 22:997–1005.

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

Lanckriet, G. R. G., Cristianini, N., Ghaoui, L. E., Bartlett, P., and Jordan, M. I. (2004b). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72.

References III

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2010). Network flow algorithms for structured sparsity. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1558–1566.

Micchelli, C. A. and Pontil, M. (2006). Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6(2):1099.

Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76.

Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2008). SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008). SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*.

Rosasco, L., Verri, A., Santoro, M., Mosci, S., and Villa, S. (2009). Iterative Projection Methods for Structured Sparsity Regularization. Technical report, CBCL-282.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge Univ Pr.

References IV

Sonnenburg, S., R
"atsch, G., Sch
"afer, C., and Sch
"olkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565.

Vishwanathan, S., sun, Z., Ampornpunt, N., and Varma, M. (2010). Multiple kernel learning and the smo algorithm. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 2361–2369.

Xu, Z., Jin, R., King, I., and Lyu, M. (2009). An extended level method for efficient multiple kernel learning. *Advances in neural information processing systems*, 21:1825–1832.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.