High-dimensional non-linear variable selection through hierarchical kernel learning

Francis Bach
Willow project, INRIA - Ecole Normale Supérieure

April 2009
Outline

- Supervised learning and regularization
  - *Kernel methods vs. sparse methods*

- MKL: Multiple kernel learning
  - *Non linear sparse methods*

- HKL: Hierarchical kernel learning
  - *Non linear variable selection*
Supervised learning and regularization

• Data: \( x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n \)

• Minimize with respect to function \( f : \mathcal{X} \to \mathcal{Y} \):

\[
\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\mu}{2} \| f \|_2^2
\]

Error on data + Regularization

Loss & function space ? Norm ?

• Two theoretical/algorithmic issues:

1. Loss
2. Function space / norm
Regularizations

- Main goal: avoid overfitting

- Two main lines of work:
  1. **Euclidean** and **Hilbertian** norms (i.e., $\ell^2$-norms)
     - Non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developed theory (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)
Regularizations

- Main goal: avoid overfitting

- Two main lines of work:

  1. **Euclidean and Hilbertian** norms (i.e., $\ell^2$-norms)
     - Non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developed theory (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)

  2. **Sparsity-inducing** norms
     - Usually restricted to linear predictors on vectors $f(x) = w^T x$
     - Main example: $\ell_1$-norm $\|w\|_1 = \sum_{i=1}^{p} |w_i|$
     - Perform model selection as well as regularization
     - Theory “in the making”
Kernel methods: regularization by $\ell^2$-norm

- Data: $x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n$, with features $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$
  - Predictor $f(x) = w^\top \Phi(x)$ linear in the features

- Optimization problem:
  $\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|^2_2$
Kernel methods: regularization by $\ell^2$-norm

- **Data:** $x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n$, with **features** $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$
  
  - Predictor $f(x) = w^\top \Phi(x)$ linear in the features

- **Optimization problem:**
  $$
  \min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|^2_2
  $$

- **Representer theorem** (Kimeldorf and Wahba, 1971): solution must be of the form $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$

  - Equivalent to solving:
    $$
    \min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\mu}{2} \alpha^\top K\alpha
    $$

  - Kernel matrix $K_{ij} = k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$
Kernel methods: regularization by $\ell^2$-norm

- Running time $O(n^2\kappa + n^3)$ where $\kappa$ complexity of one kernel evaluation (often much less) - **independent from** $p$

- **Kernel trick**: implicit mapping if $\kappa = o(p)$ by using only $k(x_i, x_j)$ instead of $\Phi(x_i)$

- Examples:
  - Polynomial kernel: $k(x, y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} = \text{polynomials}$
  - Gaussian kernel: $k(x, y) = e^{-\alpha \|x-y\|^2_2} \Rightarrow \mathcal{F} = \text{smooth functions}$
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)
Kernel methods: regularization by $\ell^2$-norm

- Running time $O(n^2\kappa + n^3)$ where $\kappa$ complexity of one kernel evaluation (often much less) - independent from $p$

- **Kernel trick**: implicit mapping if $\kappa = o(p)$ by using only $k(x_i, x_j)$ instead of $\Phi(x_i)$

- Examples:
  - Polynomial kernel: $k(x, y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} = \text{polynomials}$
  - Gaussian kernel: $k(x, y) = e^{-\alpha \|x-y\|_2^2} \Rightarrow \mathcal{F} = \text{smooth functions}$
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)

- $+$: Implicit non linearities and high-dimensionality

- $-$: Problems of interpretability, dimension too high?
\( \ell_1 \)-norm regularization (linear setting)

- Data: covariates \( x_i \in \mathbb{R}^p \), responses \( y_i \in \mathcal{Y}, i = 1, \ldots, n \)

- Minimize with respect to loadings/weights \( w \in \mathbb{R}^p \):

  \[
  \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \mu \|w\|_1
  \]

  Error on data + Regularization

- Including a constant term \( b \)? Penalizing or constraining?

- square loss \( \Rightarrow \) basis pursuit (signal processing) (Chen et al., 2001), Lasso (statistics/machine learning) (Tibshirani, 1996)
\( \ell^2 \)-norm vs. \( \ell^1 \)-norm

- \( \ell^1 \)-norms lead to interpretable models
- \( \ell^2 \)-norms can be run implicitly with very large feature spaces

**Algorithms:**
- Smooth convex optimization vs. nonsmooth convex optimization

**Theory:**
- better predictive performance?
$\ell^2$ vs. $\ell^1$ - Gaussian hare vs. Laplacian tortoise

- First-order methods (Fu, 1998; Wu and Lange, 2008)
- Homotopy methods (Markowitz, 1956; Efron et al., 2004)
1. **Consistency condition** (Zhao and Yu, 2006; Wainwright, 2006; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

\[
\| Q_{J^cJ} Q_{JJ}^{-1} \text{sign}(w_J) \|_\infty \leq 1,
\]

where \( Q = \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{p \times p} \).
Lasso - Two main recent theoretical results

1. **Consistency condition** (Zhao and Yu, 2006; Wainwright, 2006; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

\[
\| Q_{J^c} Q_J^{-1} \text{sign}(w_J) \|_\infty \leq 1,
\]

where

\[
Q = \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{p \times p}.
\]

2. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): under appropriate assumptions, consistency is possible as long as

\[
\log p = o(n)
\]
Outline

- Supervised learning and regularization
  - *Kernel methods vs. sparse methods*

- MKL: Multiple kernel learning
  - *Non linear sparse methods*

- HKL: Hierarchical kernel learning
  - *Non linear variable selection*
Multiple kernel learning (MKL)
(Lanckriet et al., 2004b; Bach et al., 2004a)

- Sparse methods are linear!

- Sparsity with non-linearities
  - replace \( f(x) = \sum_{j=1}^{p} w_j^\top x_j \) with \( x_j \in \mathbb{R} \) and \( w_j \in \mathbb{R} \)
  - by \( f(x) = \sum_{j=1}^{p} w_j^\top \Phi_j(x) \) with \( \Phi_j(x) \in \mathcal{F}_j \) an \( w_j \in \mathcal{F}_j \)

- Replace the \( \ell^1 \)-norm \( \sum_{j=1}^{p} |w_j| \) by “block” \( \ell^1 \)-norm \( \sum_{j=1}^{p} \|w_j\|_2 \)

- Remarks
  - Hilbert space extension of the group Lasso (Yuan and Lin, 2006)
  - Alternative sparsity-inducing norms (Ravikumar et al., 2008)
Multiple kernel learning (MKL)  
(Lanckriet et al., 2004b; Bach et al., 2004a)

- Multiple feature maps / kernels on \( x \in \mathcal{X} \):
  - \( p \) “feature maps” \( \Phi_j : \mathcal{X} \rightarrow \mathcal{F}_j, j = 1, \ldots, p \).
  - Minimization with respect to \( w_1 \in \mathcal{F}_1, \ldots, w_p \in \mathcal{F}_p \).
  - Predictor: \( f(x) = w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \).

\[
\begin{array}{c c c c}
\Phi_1(x)^\top & w_1 \\
\vdots & \vdots & \vdots & \vdots \\
x & \Phi_j(x)^\top & w_j & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
\vdots & \vdots & \vdots & \vdots \\
\Phi_p(x)^\top & w_p
\end{array}
\]

- Generalized additive models (Hastie and Tibshirani, 1990)
- Link between regularization and kernel matrices
Regularization for multiple features

\[
\Phi_1(x) \top w_1 \\
\vdots \quad \vdots \\
x \rightarrow \Phi_j(x) \top w_j \\
\vdots \quad \vdots \\
\Phi_p(x) \top w_p \\
\]

\[
w_1 \Phi_1(x) + \cdots + w_p \Phi_p(x)
\]

• Regularization by \( \sum_{j=1}^{p} \|w_j\|_2^2 \) is equivalent to using \( K = \sum_{j=1}^{p} K_j \)
  - Summing kernels is equivalent to concatenating feature spaces
Regularization for multiple features

\[ \Phi_1(x)^\top w_1 \]

\[ \Phi_2(x)^\top w_2 \]  
\[ \vdots \]  
\[ \Phi_p(x)^\top w_p \]

- Regularization by \( \sum_{j=1}^{p} \| w_j \|_2^2 \) is equivalent to using \( K = \sum_{j=1}^{p} K_j \)

- Regularization by \( \sum_{j=1}^{p} \| w_j \|_2 \) imposes sparsity at the group level

- **Main questions when regularizing by block \( \ell^1 \)-norm:**
  1. Algorithms (Bach et al., 2004a,b; Rakotomamonjy et al., 2008)
  2. Analysis of sparsity inducing properties (Bach, 2008a)
  3. Does it correspond to a specific combination of kernels?
• **Proposition** (Lanckriet et al., 2004, Bach et al., 2005, Micchelli and Pontil, 2005):

\[
G(K) = \min_{w \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|^2
\]

\[
= \max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^{n} \ell_i^*(\mu \alpha_i) - \frac{\mu}{2} \alpha^\top K \alpha
\]

is a convex function of the kernel matrix \( K \)

• Theoretical learning **bounds** (Lanckriet et al., 2004, Srebro and Ben-David, 2006)
Equivalence with kernel learning (Bach et al., 2004a)

- Block $\ell^1$-norm problem:

$$\sum_{i=1}^{n} \ell(y_i, w_1^\top \Phi_1(x_i) + \cdots + w_p^\top \Phi_p(x_i)) + \frac{\mu}{2} (\|w_1\|_2 + \cdots + \|w_p\|_2)^2$$

- Proposition: Block $\ell^1$-norm regularization is equivalent to minimizing with respect to $\eta$ the optimal value $G(\sum_{j=1}^{p} \eta_j K_j)$

- (sparse) weights $\eta$ obtained from optimality conditions

- dual parameters $\alpha$ optimal for $K = \sum_{j=1}^{p} \eta_j K_j$

- Single optimization problem for learning both $\eta$ and $\alpha$
Analysis of MKL as non parametric group Lasso

- Assume $p$ Hilbert spaces $\mathcal{F}_i, i = 1, \ldots, p$ on $p$ different input spaces

\[
\min_{f_1 \in \mathcal{F}_1, \ldots, f_p \in \mathcal{F}_p} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} f_j(x_{ji}) \right)^2 + \frac{\mu n}{2} \left( \sum_{j=1}^{p} \|f_j\| \right)^2.
\]

NB: $f_j(x_{ji}) = f_j^\top \Phi_j(x_{ji})$

- Sparse generalized additive models (Hastie and Tibshirani, 1990, Ravikumar et al., 2007)

- **Algorithms**: use parametrization with $\alpha$

- **Analysis**: Do not use $\alpha \Rightarrow$ use covariance operators (i.e., stay in the primal/input space)
Covariance operators

• Single random variable $X$: $\Sigma_{XX}$ is a bounded linear operator from $\mathcal{F}$ to $\mathcal{F}$ such that for all $(f, g) \in \mathcal{F} \times \mathcal{F}$,

$$\langle f, \Sigma_{XX} g \rangle = \text{cov}(f(X), g(X))$$

Under minor assumptions, the operator $\Sigma_{XX}$ is auto-adjoint, non-negative and Hilbert-Schmidt

• Tool of choice for the analysis of least-squares non parametric methods (Blanchard, 2006, Fukumizu et al., 2005, 2006, Gretton et al., 2006, Harchaoui et al., 2007, 2008, etc.)

  – Natural empirical estimate $\langle f, \hat{\Sigma}_{XX} g \rangle = \hat{\text{cov}}(f(X), g(X))$ converges in probability to $\Sigma_{XX}$ in HS norm.
Cross-covariance operators

- Several random variables: cross-covariance operators $\Sigma_{X_iX_j}$ from $\mathcal{F}_j$ to $\mathcal{F}_i$ such that $\forall (f_i, f_j) \in \mathcal{F}_i \times \mathcal{F}_j$,

$$\langle f_i, \Sigma_{X_iX_j} f_j \rangle = \text{cov}(f_i(X_i), f_j(X_j))$$

- Similar convergence properties of empirical estimates

- Joint covariance operator $\Sigma_{XX}$ defined by blocks

- We can define the bounded correlation operators through

$$\Sigma_{X_iX_j} = \Sigma^{1/2}_{X_iX_i} C_{X_iX_j} \Sigma^{1/2}_{X_jX_j}$$

- NB: the joint covariance operator is never invertible, but the correlation operator may be
Analysis of MKL as non parametric group Lasso

• Assumptions

1. **Generalized additive model**: There exists functions \( f = (f_1, \ldots, f_p) \in F = F_1 \times \cdots \times F_p \) such that

   \[
   Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon
   \]

2. **Compacity and invertibility**: All cross-correlation operators are compact and the joint correlation operator is invertible.

3. **Smoothness of predictors**: \( \forall j \in \{1, \ldots, p\}, \sum_{X_j X_j}^{-1/2} f_j \in F_j \)
Compacity and invertibility of joint correlation operator

• Sufficient condition for compacity when distributions have densities:

\[
\mathbb{E} \left\{ \frac{p_{X_i X_j}(x_i, x_j)}{p_{X_i}(x_i)p_{X_j}(x_j)} - 1 \right\} < \infty.
\]

– Dependence between variables is not too strong

• Sufficient condition for invertibility: no exact correlation using functions in the RKHS.

– Empty *concurvity* space assumption (Hastie and Tibshirani, 1990)
Group lasso - Consistency conditions

• Strict condition

$$\max_{i \in J^c} \left\| \Sigma_{X_iX_i}^{1/2} C_{X_iX_j} C_{X_jX_j}^{-1} (\Sigma_{X_jX_j}^{-1/2} f_j / \|f_j\|)_{j \in J} \right\| < 1$$

• Weak condition

$$\max_{i \in J^c} \left\| \Sigma_{X_iX_i}^{1/2} C_{X_iX_j} C_{X_jX_j}^{-1} (\Sigma_{X_jX_j}^{-1/2} f_j / \|f_j\|)_{j \in J} \right\| \leq 1$$

• **Theorem 1**: Strict condition is **sufficient** for joint regular and sparsity consistency of the lasso.

• **Theorem 2**: Weak condition is **necessary**.
Group lasso - Consistency conditions

- **Strict condition**

\[
\max_{i \in J^c} \left\| \sum_{X_iX_i}^{1/2} C_{X_iX_J} C_{X_JX_J}^{-1} (\Sigma_{X_jX_j}^{-1/2} f_j / \| f_j \|) \right\|_j < 1
\]

- **Weak condition**

\[
\max_{i \in J^c} \left\| \sum_{X_iX_i}^{1/2} C_{X_iX_J} C_{X_JX_J}^{-1} (\Sigma_{X_jX_j}^{-1/2} f_j / \| f_j \|) \right\|_j \leq 1
\]

- **Theorem 1:** Strict condition is **sufficient** for joint regular and sparsity consistency of the lasso.

- **Theorem 2:** Weak condition is **necessary.**

- **Asymptotic results!**
Applications

• Several applications
  – Bioinformatics (Lanckriet et al., 2004a)
  – Image annotation (Harchaoui and Bach, 2007; Varma and Ray, 2007; Bosch et al., 2008)

• Two potential uses
  – Fusion of heterogeneous data sources
  – Learning hyperparameters
  – Sparsity in non-linear settings
Caltech101 database (Fei-Fei et al., 2006)
# Kernel combination for Caltech101 (Varma and Ray, 2007)

## Classification accuracies

<table>
<thead>
<tr>
<th>Feature</th>
<th>1-NN</th>
<th>SVM (1 vs. 1)</th>
<th>SVM (1 vs. rest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape GB1</td>
<td>39.67 ± 1.02</td>
<td>57.33 ± 0.94</td>
<td>62.98 ± 0.70</td>
</tr>
<tr>
<td>Shape GB2</td>
<td>45.23 ± 0.96</td>
<td>59.30 ± 1.00</td>
<td>61.53 ± 0.57</td>
</tr>
<tr>
<td>Self Similarity</td>
<td>40.09 ± 0.98</td>
<td>55.10 ± 1.05</td>
<td>60.83 ± 0.84</td>
</tr>
<tr>
<td>PHOG 180</td>
<td>32.01 ± 0.89</td>
<td>48.83 ± 0.78</td>
<td>49.93 ± 0.52</td>
</tr>
<tr>
<td>PHOG 360</td>
<td>31.17 ± 0.98</td>
<td>50.63 ± 0.88</td>
<td>52.44 ± 0.85</td>
</tr>
<tr>
<td>PHOWColour</td>
<td>32.79 ± 0.92</td>
<td>40.84 ± 0.78</td>
<td>43.44 ± 1.46</td>
</tr>
<tr>
<td>PHOWGray</td>
<td>42.08 ± 0.81</td>
<td>52.83 ± 1.00</td>
<td>57.00 ± 0.30</td>
</tr>
<tr>
<td><strong>MKL Block ℓ¹</strong></td>
<td><strong>77.72 ± 0.94</strong></td>
<td><strong>83.78 ± 0.39</strong></td>
<td></td>
</tr>
<tr>
<td>(Varma and Ray, 2007)</td>
<td><strong>81.54 ± 1.08</strong></td>
<td><strong>89.56 ± 0.59</strong></td>
<td></td>
</tr>
</tbody>
</table>

- See also Bosch et al. (2008)
Outline

• Supervised learning and regularization
  – *Kernel methods vs. sparse methods*

• MKL: Multiple kernel learning
  – *Non linear sparse methods*

• HKL: Hierarchical kernel learning
  – *Non linear variable selection*
Lasso - Two main recent theoretical results

1. **Consistency condition**

2. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): under appropriate assumptions, consistency is possible as long as

\[ \log p = o(n) \]
Lasso - Two main recent theoretical results

1. **Consistency condition**

2. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): under appropriate assumptions, consistency is possible as long as

\[ \log p = o(n) \]

- Question: is it possible to build a sparse algorithm that can learn from more than \(10^{80}\) features?
Lasso - Two main recent theoretical results

1. **Consistency condition**

2. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): under appropriate assumptions, consistency is possible as long as

   \[ \log p = o(n) \]

- Question: is it possible to build a sparse algorithm that can learn from more than \(10^{80}\) features?

  - Some type of recursivity/factorization is needed!
Hierarchical kernel learning (Bach, 2008b)

- Many kernels can be decomposed as a sum of many “small” kernels
  \[ k(x, x') = \sum_{v \in V} k_v(x, x') \]

- Example with \( x = (x_1, \ldots, x_q) \in \mathbb{R}^q \) (⇒ non linear variable selection)
  - 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)

- **Goal**: learning sparse combination
  \[ \sum_{v \in V} \eta_v k_v(x, x') \]
Restricting the set of active kernels

• With flat structure
  – Consider block $\ell^1$-norm: $\sum_{v \in V} d_v \|w_v\|_2$
  – cannot avoid being linear in $p = \#(V)$

• Using the structure of the small kernels
  – for computational reasons
  – to allow more irrelevant variables
Restricting the set of active kernels

- $V$ is endowed with a directed acyclic graph (DAG) structure: select a kernel only after all of its ancestors have been selected

- Gaussian kernels: $V = \text{power set of } \{1, \ldots, q\}$ with inclusion DAG
  - Select a subset only after all its subsets have been selected
DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  $D(v)$ is the set of descendants of $v \in V$:
  $$\sum_{v \in V} d_v \|w_{D(v)}\|_2 = \sum_{v \in V} d_v \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If $v$ is selected, so are all its ancestors
DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  - \( D(v) \) is the set of descendants of \( v \in V \):
    \[
    \sum_{v \in V} d_v \| w_{D(v)} \|_2 = \sum_{v \in V} d_v \left( \sum_{t \in D(v)} \| w_t \|_2^2 \right)^{1/2}
    \]

- Main property: If \( v \) is selected, so are all its ancestors

- Questions:
  - polynomial-time algorithm for this norm?
  - necessary/sufficient conditions for consistent kernel selection?
  - Scaling between \( p, q, n \) for consistency?
  - Applications to variable selection or other kernels?
Active set algorithm for sparse problems

- First assume that the set $J$ of active kernels is known
  - If $J$ is small, solving the reduced problem is easy
  - Simply need to check if the solution is optimal for the full problem
    * If yes, the solution is found
    * If not, add violating variables to the reduced problem
Active set algorithm for sparse problems

• First assume that the set $J$ of active kernels is known
  – If $J$ is small, solving the reduced problem is easy
  – Simply need to check if the solution is optimal for the full problem
    * If yes, the solution is found
    * If not, add violating variables to the reduced problem

• Technical issue: computing approximate necessary and sufficient conditions in polynomial time in the out-degree of the DAG
  – NB: with flat structure, this is linear in $p = \#(V)$

• Active set algorithm: start with the roots of the DAG and grow
  – Running time polynomial in the number of selected kernels
Because of the selection constraints, getting the exact sparse model is not possible in general.

May only estimate the *hull* of the relevant kernels.

Necessary and sufficient conditions can be derived.
Scaling between $p$, $q$, $n$

$n = \text{number of observations}$
$q = \text{maximum out degree in the DAG}$
$p = \text{number of vertices in the DAG}$

- **Theorem**: Assume consistency condition satisfied and Gaussian noise and $\lambda = c_1 A \sigma \left( \frac{\log q}{n} \right)^{1/2}$, with $A \in \left[ 1, \left( \frac{n}{\log q} \right)^{1/2} \right]$; the probability of incorrect hull selection is upper-bounded by

$$
c_2 \exp \left( -\frac{c_3 n}{\sigma^2} \right) + \exp \left( -c_4 A^2 \log q \right).$$
Scaling between $p, q, n$

- $n = \text{number of observations}$
- $q = \text{maximum out degree in the DAG}$
- $p = \text{number of vertices in the DAG}$

**Theorem:** Assume consistency condition satisfied and Gaussian noise and $\lambda = c_1 A \sigma \left( \frac{\log q}{n} \right)^{1/2}$, with $A \in \left[ 1, \left( \frac{n}{\log q} \right)^{1/2} \right]$; the probability of incorrect hull selection is upper-bounded by

$$c_2 \exp \left( -\frac{c_3 n}{\sigma^2} \right) + \exp \left( -c_4 A^2 \log q \right).$$

**Unstructured case:** $q = p \Rightarrow n \approx \log p$

**Power set of $q$ elements:** $q = \log p \Rightarrow n \approx \log \log p = \log q$
Comparing norms on synthetic example

- Sparse non linear problem with decomposed Gaussian/ANOVA kernels
  - Left: original data (generated by sparse multiivariate polynomial)
  - Right: rotated data (sparsity non expected)

![Graph of test set error vs \( \log_2(p) \) for HKL, greedy, and L2 norms under sparsity expected and not expected conditions.](image)

* sparsity is expected
* sparsity is not expected
## Mean-square errors (regression)

<table>
<thead>
<tr>
<th>dataset</th>
<th>n</th>
<th>p</th>
<th>k</th>
<th>#(V)</th>
<th>L2</th>
<th>greedy</th>
<th>MKL</th>
<th>HKL</th>
</tr>
</thead>
<tbody>
<tr>
<td>abalone</td>
<td>4177</td>
<td>10</td>
<td>pol4</td>
<td>(\approx 10^7)</td>
<td>44.2±1.3</td>
<td>43.9±1.4</td>
<td>44.5±1.1</td>
<td>43.3±1.0</td>
</tr>
<tr>
<td>abalone</td>
<td>4177</td>
<td>10</td>
<td>rbf</td>
<td>(\approx 10^{10})</td>
<td>43.0±0.9</td>
<td>45.0±1.7</td>
<td>43.7±1.0</td>
<td>43.0±1.1</td>
</tr>
<tr>
<td>boston</td>
<td>506</td>
<td>13</td>
<td>pol4</td>
<td>(\approx 10^9)</td>
<td>17.1±3.6</td>
<td>24.7±10.8</td>
<td>22.2±2.2</td>
<td>18.1±3.8</td>
</tr>
<tr>
<td>boston</td>
<td>506</td>
<td>13</td>
<td>rbf</td>
<td>(\approx 10^{12})</td>
<td>16.4±4.0</td>
<td>32.4±8.2</td>
<td>20.7±2.1</td>
<td>17.1±4.7</td>
</tr>
<tr>
<td>pumadyn-32fh</td>
<td>8192</td>
<td>32</td>
<td>pol4</td>
<td>(\approx 10^{22})</td>
<td>57.3±0.7</td>
<td>56.4±0.8</td>
<td>56.4±0.7</td>
<td>56.4±0.8</td>
</tr>
<tr>
<td>pumadyn-32fh</td>
<td>8192</td>
<td>32</td>
<td>rbf</td>
<td>(\approx 10^{31})</td>
<td>57.7±0.6</td>
<td>72.2±22.5</td>
<td>56.5±0.8</td>
<td>55.7±0.7</td>
</tr>
<tr>
<td>pumadyn-32fm</td>
<td>8192</td>
<td>32</td>
<td>pol4</td>
<td>(\approx 10^{22})</td>
<td>6.9±0.1</td>
<td>6.4±1.6</td>
<td>7.0±0.1</td>
<td>3.1±0.0</td>
</tr>
<tr>
<td>pumadyn-32fm</td>
<td>8192</td>
<td>32</td>
<td>rbf</td>
<td>(\approx 10^{31})</td>
<td>5.0±0.1</td>
<td>46.2±51.6</td>
<td>7.1±0.1</td>
<td>3.4±0.0</td>
</tr>
<tr>
<td>pumadyn-32nh</td>
<td>8192</td>
<td>32</td>
<td>pol4</td>
<td>(\approx 10^{22})</td>
<td>84.2±1.3</td>
<td>73.3±25.4</td>
<td>83.6±1.3</td>
<td>36.7±0.4</td>
</tr>
<tr>
<td>pumadyn-32nh</td>
<td>8192</td>
<td>32</td>
<td>rbf</td>
<td>(\approx 10^{31})</td>
<td>56.5±1.1</td>
<td>81.3±25.0</td>
<td>83.7±1.3</td>
<td>35.5±0.5</td>
</tr>
<tr>
<td>pumadyn-32nm</td>
<td>8192</td>
<td>32</td>
<td>pol4</td>
<td>(\approx 10^{22})</td>
<td>60.1±1.9</td>
<td>69.9±32.8</td>
<td>77.5±0.9</td>
<td>5.5±0.1</td>
</tr>
<tr>
<td>pumadyn-32nm</td>
<td>8192</td>
<td>32</td>
<td>rbf</td>
<td>(\approx 10^{31})</td>
<td>15.7±0.4</td>
<td>67.3±42.4</td>
<td>77.6±0.9</td>
<td>7.2±0.1</td>
</tr>
</tbody>
</table>
Extensions to other kernels

- Extension to graph kernels, string kernels, pyramid match kernels
- Exploring large feature spaces with structured sparsity-inducing norms
  - Interpretable models
- Other structures than hierarchies or DAGs
Summary

• Supervised learning and regularization
  – *Kernel methods vs. sparse methods*

• MKL: Multiple kernel learning
  – *Non linear sparse methods*

• HKL: Hierarchical kernel learning
  – *Non linear variable selection*
Extensions - Conclusion

• Further/current work
  – Consistency of non linear variable selection
  – Universal consistency
  – Algorithms
  – norm design, norms on matrices
  – General overlapping groups (Jenatton et al., 2009)
  – Computer vision and bioinformatics

• Sparsity and non-linearity are not incompatible
  – Incorporate structure into the learning problem
References


