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: $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, \dots, n$
- Minimize with respect to function $f : \mathcal{X} \to \mathcal{Y}$:

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

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., ℓ^2-norms)
     - Non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped 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., ℓ^2-norms)
   - Non linear predictors
   - Non parametric supervised learning and kernel methods
   - Well developped 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 $f(x) = w^\top x$
   - Main example: ℓ_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 ℓ^2-norm

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

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

Kernel methods: regularization by ℓ^2-norm

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

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

- **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: $$\boxed{\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\mu}{2} \alpha^\top K \alpha}$$
  - Kernel matrix $K_{ij} = k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

Kernel methods: regularization by ℓ^2-norm

- Running time $O(n^2\kappa + n^3)$ where κ 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: $k(x,y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} =$ polynomials
  - Gaussian kernel: $k(x,y) = e^{-\alpha\|x-y\|_2^2} \Rightarrow \mathcal{F} =$ smooth functions
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)

Kernel methods: regularization by ℓ^2-norm

- Running time $O(n^2\kappa + n^3)$ where κ 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: $k(x, y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} =$ polynomials
  - Gaussian kernel: $k(x, y) = e^{-\alpha\|x-y\|_2^2} \Rightarrow \mathcal{F} =$ smooth functions
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)

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

- $-$: Problems of interpretability, dimension too high?

ℓ_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 $w \in \mathbb{R}^p$:

$$\underbrace{\sum_{i=1}^{n} \ell(y_i, w^\top x_i)}_{\text{Error on data}} \quad + \quad \underbrace{\mu \|w\|_1}_{\text{Regularization}}$$

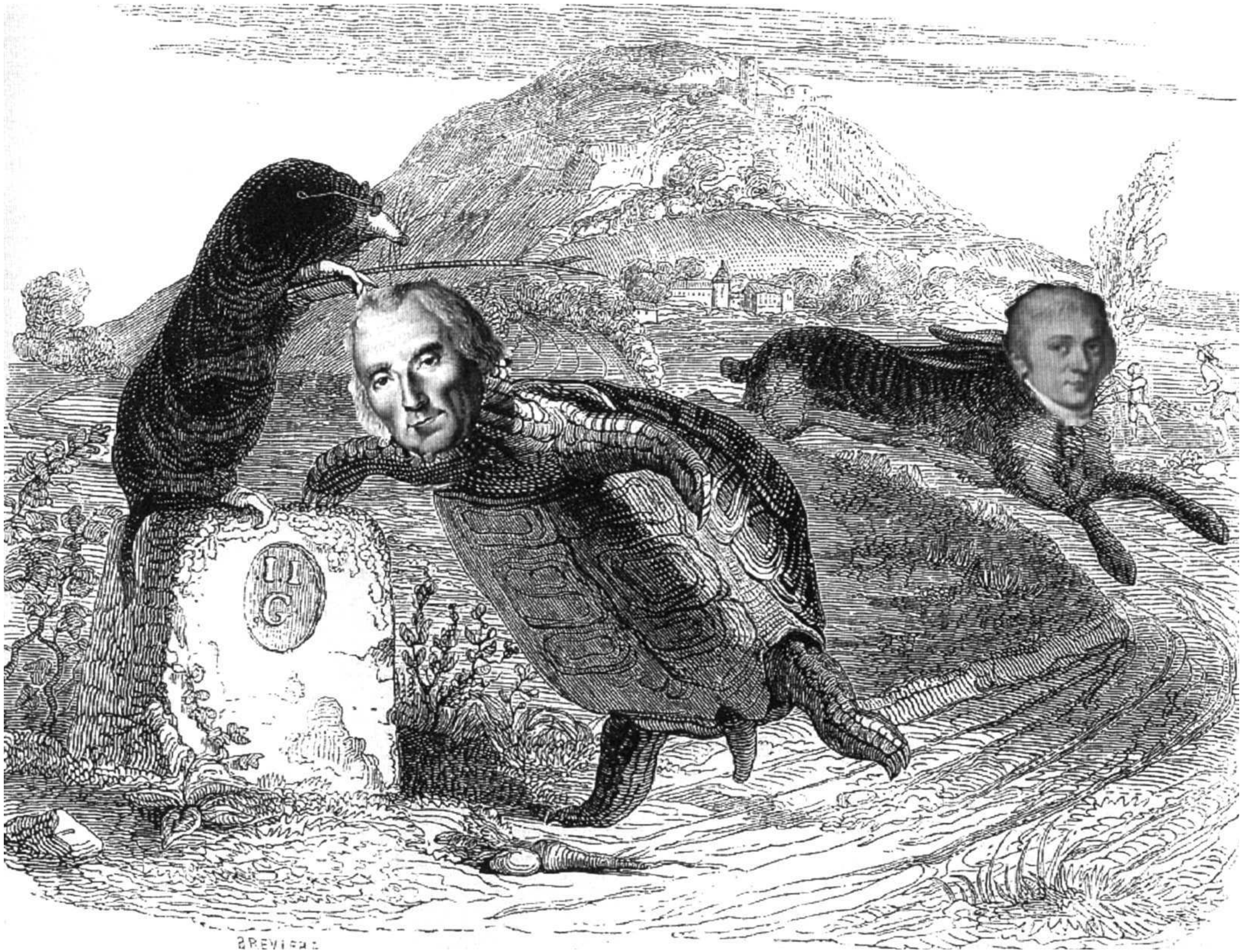
- Including a constant term b? Penalizing or constraining?

- square loss $\Rightarrow$ basis pursuit (signal processing) (Chen et al., 2001), Lasso (statistics/machine learning) (Tibshirani, 1996)

ℓ^2-norm vs. ℓ^1-norm

- ℓ^1-norms lead to interpretable models
- ℓ^2-norms can be run implicitly with very large feature spaces
- **Algorithms**:
  - Smooth convex optimization vs. nonsmooth convex optimization
- **Theory**:
  - better predictive performance?

ℓ^2 vs. ℓ^1 - Gaussian hare vs. Laplacian tortoise



- First-order methods (Fu, 1998; Wu and Lange, 2008)
- Homotopy methods (Markowitz, 1956; Efron et al., 2004)

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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathrm{sign}(\mathbf{w}_\mathbf{J})\|_\infty \leqslant 1,$$

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

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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathrm{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1,$$

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

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 ℓ^1-norm $\sum_{j=1}^{p} |w_j|$ by "block" ℓ^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" $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j$, $j = 1, \dots, p$.
  - Minimization with respect to $w_1 \in \mathcal{F}_1, \dots, w_p \in \mathcal{F}_p$
  - Predictor: $f(x) = w_1{}^\top \Phi_1(x) + \cdots + w_p{}^\top \Phi_p(x)$

$$
\begin{array}{ccccccc}
 & & \Phi_1(x)^\top & w_1 & & & \\
 & \nearrow & \vdots & \vdots & \searrow & & \\
x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
 & \searrow & \vdots & \vdots & \nearrow & & \\
 & & \Phi_p(x)^\top & w_p & & &
\end{array}
$$

  - Generalized additive models (Hastie and Tibshirani, 1990)
  - **Link between regularization and kernel matrices**

Regularization for multiple features

$$\begin{array}{ccccccc}
 & \nearrow & \Phi_1(x)^\top & w_1 & \searrow & \\
 & & \vdots & \vdots & & \\
x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
 & & \vdots & \vdots & & \\
 & \searrow & \Phi_p(x)^\top & w_p & \nearrow &
\end{array}$$

- 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

$$x \begin{array}{c} \nearrow \\ \\ \longrightarrow \\ \\ \searrow \end{array} \begin{array}{cc} \Phi_1(x)^\top & w_1 \\ \vdots & \vdots \\ \Phi_j(x)^\top & w_j \\ \vdots & \vdots \\ \Phi_p(x)^\top & w_p \end{array} \begin{array}{c} \searrow \\ \\ \longrightarrow \\ \\ \nearrow \end{array} w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x)$$

- 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 ℓ^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?

General kernel learning

- **Proposition** (Lanckriet et al, 2004, Bach et al., 2005, Micchelli and Pontil, 2005):

$$
\begin{aligned}
G(K) &= \min_{w \in \mathcal{F}} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|_2^2 \\
&= \max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^n \ell_i^*(\mu \alpha_i) - \frac{\mu}{2} \alpha^\top K \alpha
\end{aligned}
$$

  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 ℓ^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 ℓ^1-norm regularization is equivalent to minimizing with respect to η the optimal value $G(\sum_{j=1}^{p} \eta_j K_j)$
- (sparse) weights η obtained from optimality conditions
- dual parameters α optimal for $K = \sum_{j=1}^{p} \eta_j K_j$,
- **Single optimization problem for learning both η and α**

Analysis of MKL as non parametric group Lasso

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

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

  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 α

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

Covariance operators

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

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

  Under minor assumptions, the operator Σ_{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 $\langle f, \hat{\Sigma}_{XX} g \rangle = \widehat{\mathrm{cov}}(f(X), g(X))$ converges in probability to Σ_{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 $\forall (f_i, f_j) \in \mathcal{F}_i \times \mathcal{F}_j$,

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

- Similar convergence properties of empirical estimates

- Joint covariance operator Σ_{XX} defined by blocks

- We can define the bounded *correlation* operators through

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

- 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 $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_p) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_p$ such that

$$Y = \sum_{j=1}^{p} \mathbf{f}_j(X_j) + \varepsilon$$

2. **Compacity and invertibility** : All cross-correlation operators are compact and the joint correlation operator is invertible.
3. **Smoothness of predictors**: $\forall j \in \{1, \ldots, p\}$, $\Sigma_{X_j X_j}^{-1/2} \mathbf{f}_j \in \mathcal{F}_j$

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

$$\max_{i \in \mathbf{J}^c} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} (\Sigma_{X_j X_j}^{-1/2} \mathbf{f}_j / \|\mathbf{f}_j\|)_{j \in \mathbf{J}} \right\| < 1$$

- Weak condition

$$\max_{i \in \mathbf{J}^c} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} (\Sigma_{X_j X_j}^{-1/2} \mathbf{f}_j / \|\mathbf{f}_j\|)_{j \in \mathbf{J}} \right\| \leqslant 1$$

- **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

$$\max_{i \in \mathbf{J}^c} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} (\Sigma_{X_j X_j}^{-1/2} \mathbf{f}_j / \|\mathbf{f}_j\|)_{j \in \mathbf{J}} \right\| < 1$$

- Weak condition

$$\max_{i \in \mathbf{J}^c} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} (\Sigma_{X_j X_j}^{-1/2} \mathbf{f}_j / \|\mathbf{f}_j\|)_{j \in \mathbf{J}} \right\| \leqslant 1$$

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

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

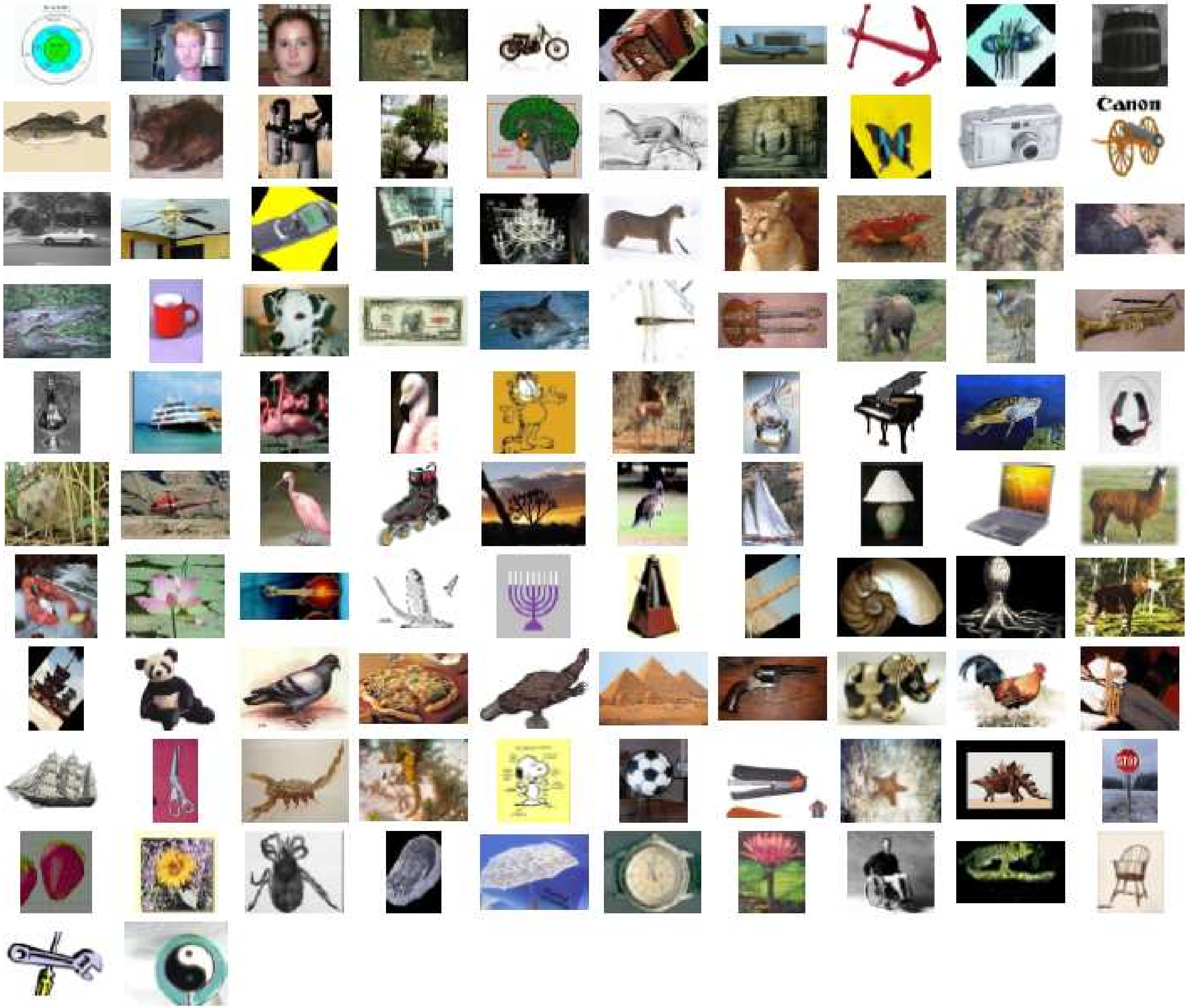
- **Asymptotic results!**

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

	1- NN	SVM (1 vs. 1)	SVM (1 vs. rest)
Shape GB1	39.67 ± 1.02	57.33 ± 0.94	62.98 ± 0.70
Shape GB2	45.23 ± 0.96	59.30 ± 1.00	61.53 ± 0.57
Self Similarity	40.09 ± 0.98	55.10 ± 1.05	60.83 ± 0.84
PHOG 180	32.01 ± 0.89	48.83 ± 0.78	49.93 ± 0.52
PHOG 360	31.17 ± 0.98	50.63 ± 0.88	52.44 ± 0.85
PHOWColour	32.79 ± 0.92	40.84 ± 0.78	43.44 ± 1.46
PHOWGray	42.08 ± 0.81	52.83 ± 1.00	57.00 ± 0.30
MKL Block ℓ^1		$\mathbf{77.72 \pm 0.94}$	$\mathbf{83.78 \pm 0.39}$
(Varma and Ray, 2007)		$\mathbf{81.54 \pm 1.08}$	$\mathbf{89.56 \pm 0.59}$

- 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 $x = (x_1, \dots, x_q) \in \mathbb{R}^q$ ($\Rightarrow$ non linear variable selection)
  - Gaussian/ANOVA kernels: $p = \#(V) = 2^q$

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

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

- **Goal**: learning sparse combination $\sum_{v \in V} \eta_v k_v(x, x')$

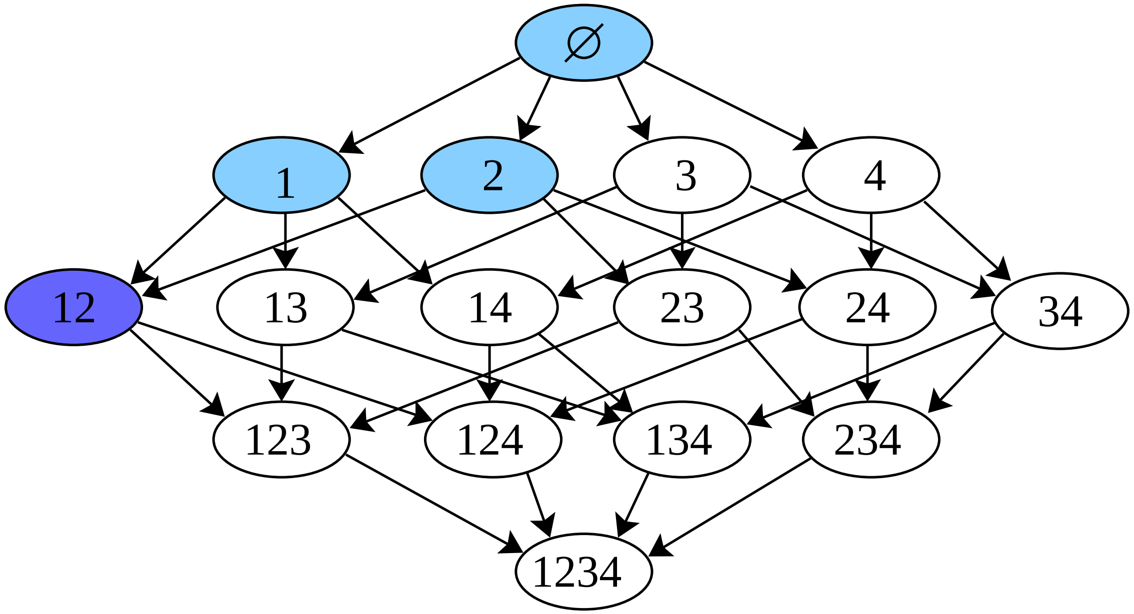
Restricting the set of active kernels

- With flat structure
  - Consider block ℓ^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: $V =$ power set of $\{1, \ldots, q\}$ with inclusion DAG
  – Select a subset only after all its subsets have been selected

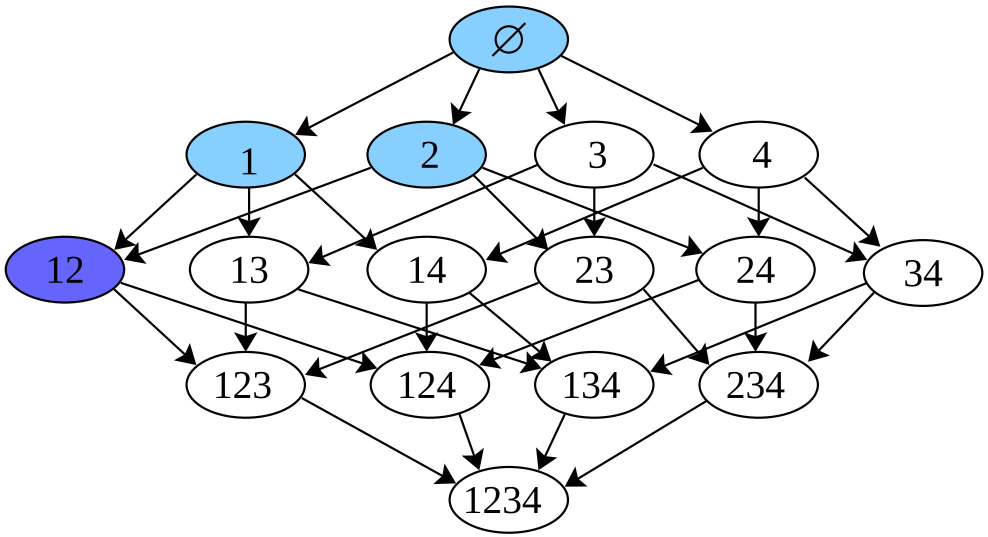


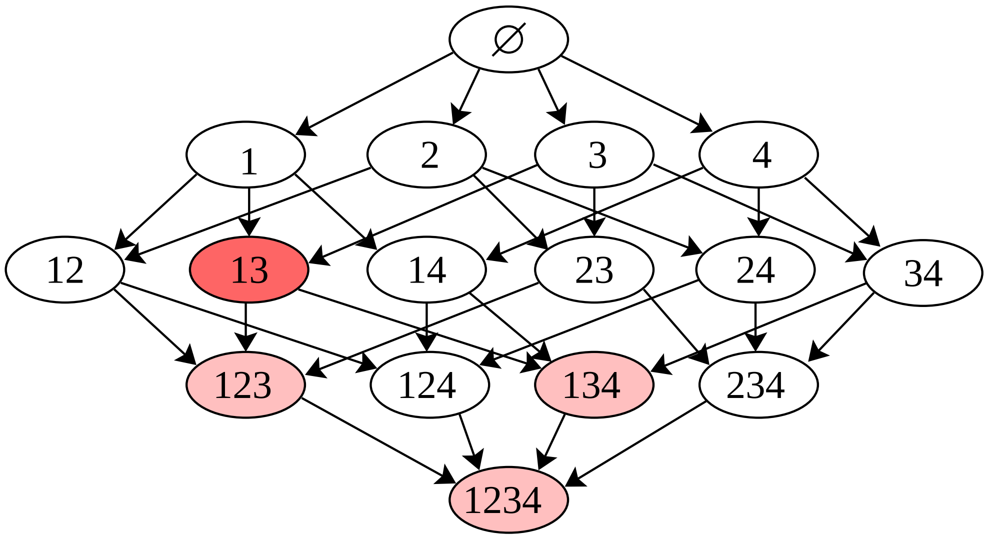
DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  - $\mathrm{D}(v)$ is the set of descendants of $v \in V$:

$$\sum_{v \in V} d_v \|w_{\mathrm{D}(v)}\|_2 = \sum_{v \in V} d_v \left(\sum_{t \in \mathrm{D}(v)} \|w_t\|_2^2 \right)^{1/2}$$

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





DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  - $\mathrm{D}(v)$ is the set of descendants of $v \in V$:

$$\sum_{v \in V} d_v \|w_{\mathrm{D}(v)}\|_2 = \sum_{v \in V} d_v \left(\sum_{t \in \mathrm{D}(v)} \|w_t\|_2^2 \right)^{1/2}$$

- 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

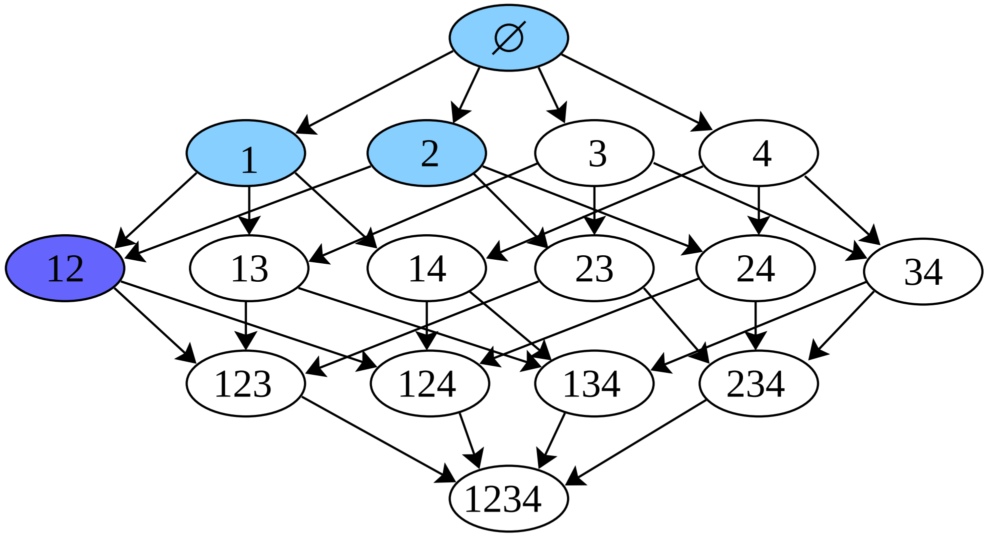
Active set algorithm for sparse problems

- 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

Consistency of kernel selection (Bach, 2008b)



- 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

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

- **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

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

- **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

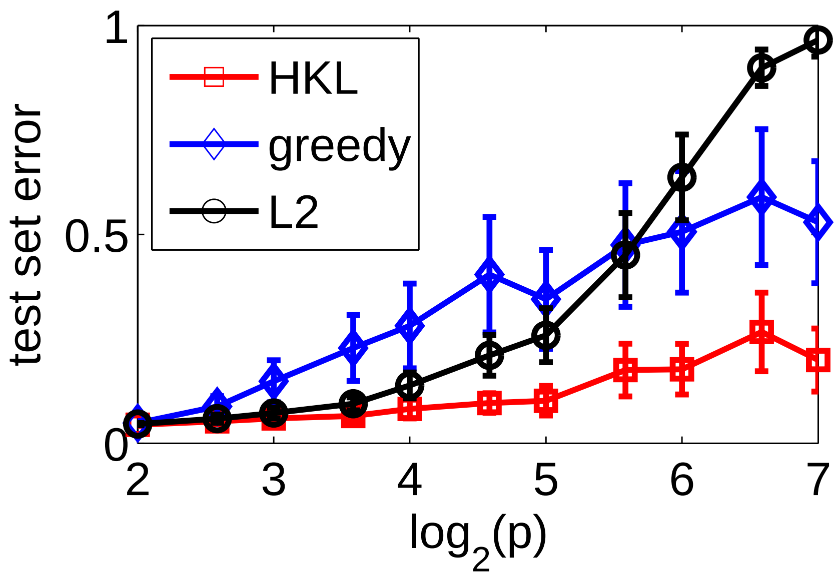
$$c_2 \exp\left(-\frac{c_3 n}{\sigma^2}\right) + \exp\left(-c_4 A^2 \log q\right).$$

- Unstructured case: $q = p \Rightarrow \boxed{n \approx \log p}$

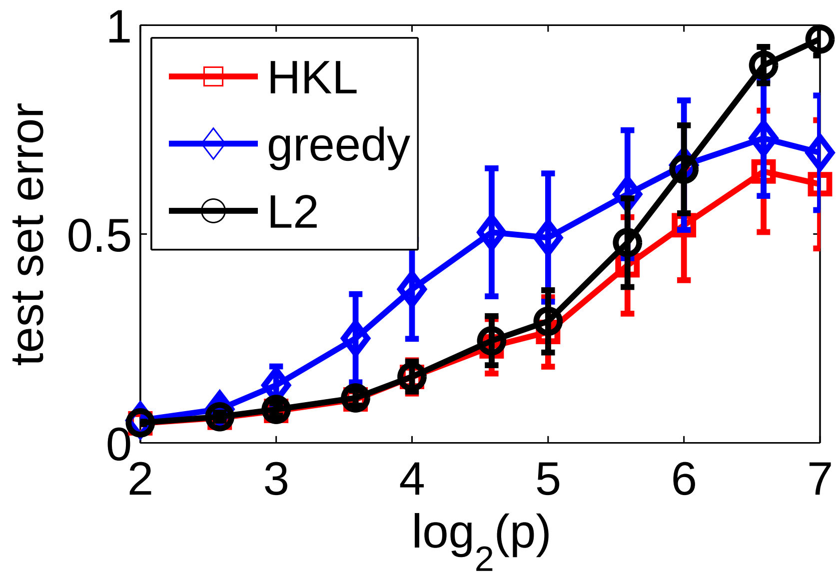
- Power set of q elements: $q = \log p \Rightarrow \boxed{n \approx \log\log p = \log q}$

Comparing norms on synthetic example

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



sparsity is expected



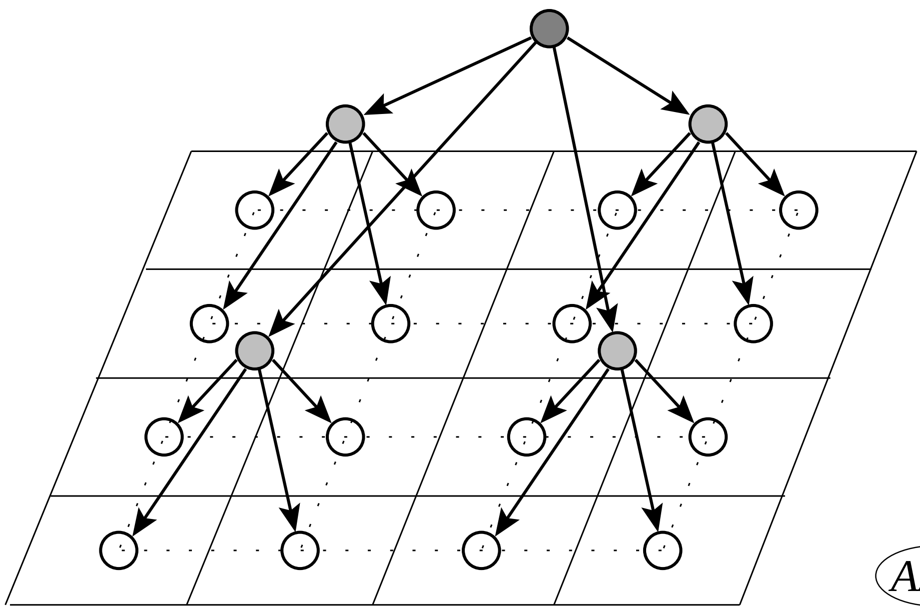
sparsity is not expected

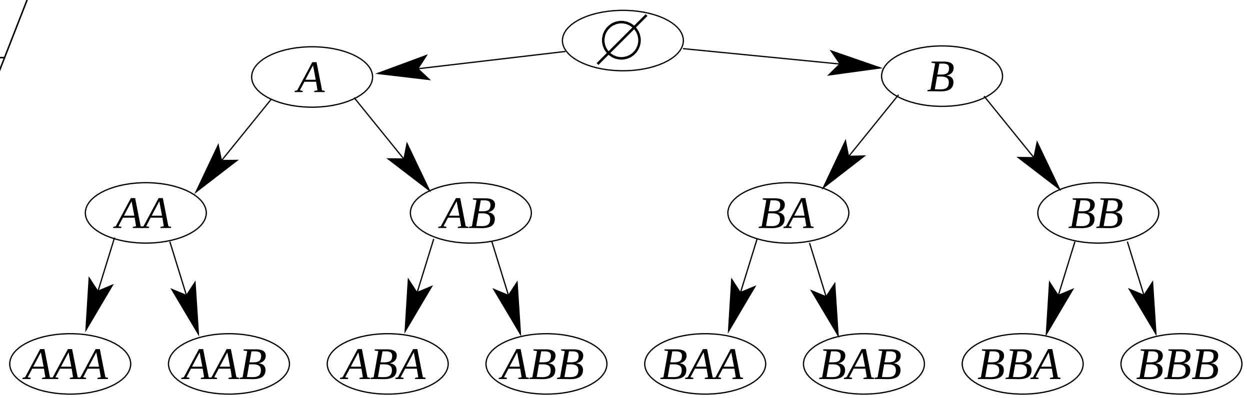
Mean-square errors (regression)

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

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

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, pages 1179–1225, 2008a.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Adv. NIPS*, 2008b.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004a.

F. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems 17*, 2004b.

P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. To appear.

A. Bosch, Zisserman A., and X. Munoz. Image classification using rois and multiple kernel learning. *International Journal of Computer Vision*, 2008. submitted.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001. ISSN 0036-1445.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32:407, 2004.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models for 101 object categories. *Computer Vision and Image Understanding*, 2006.

W. Fu. Penalized regressions: the bridge vs. the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998).

Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009. Submitted.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.

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

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

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

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2, 2008.

H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3:111–133, 1956.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Stat.*, 2009. to appear.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, to appear, 2008.

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

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Camb. U. P., 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. ICCV*, 2007.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical Report 709, Dpt. of Statistics, UC Berkeley, 2006.

T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, 2(1):224–244, 2008.

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

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

P. Zhao and B. Yu. On model selection consistency of Lasso. *JMLR*, 7:2541–2563, 2006.

H. Zou. The adaptive Lasso and its oracle properties. *JASA*, 101:1418–1429, 2006.