Sparse methods for machine learning
Theory and algorithms

Francis Bach **Guillaume Obozinski**

Sierra project-team, INRIA - Ecole Normale Supérieure

Mascot-Num, March 2012

Supervised learning and regularization

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

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

Error on data + Regularization

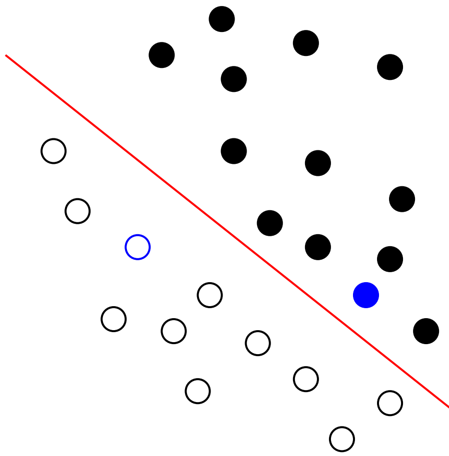
Loss & function space ? Norm ?

- Two theoretical/algorithmic issues:

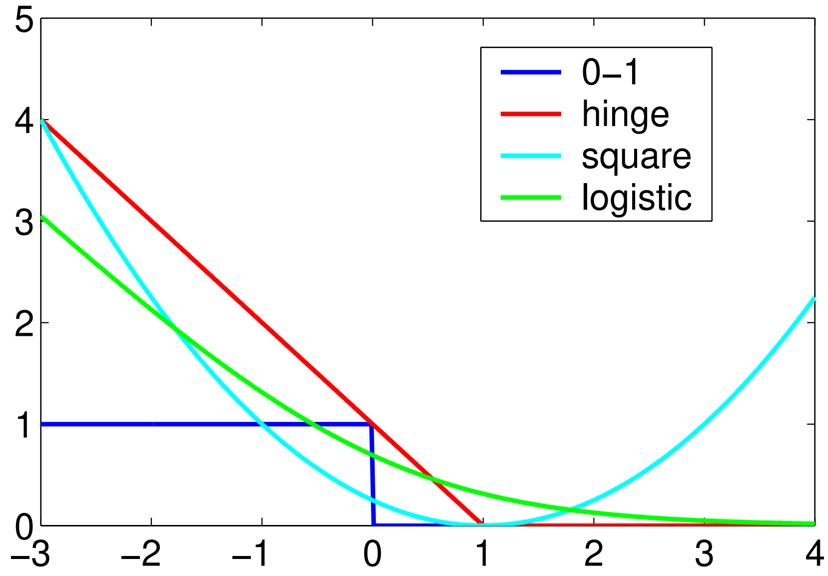
1. Loss
2. **Function space / norm**

Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = f(x)$, quadratic cost $\ell(y, f) = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - f)^2$



- **Classification** : $y \in \{-1, 1\}$ prediction $\hat{y} = \text{sign}(f(x))$
  - loss of the form $\ell(y, f) = \ell(yf)$
  - "True" cost: $\ell(yf) = 1_{yf<0}$
  - Usual convex costs:



Regularizations

- **Main goal: avoid overfitting**

- **Two main lines of work**:

  1. Euclidean and Hilbertian norms (i.e., ℓ_2-norms)
     - Possibility of non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped theory and algorithms (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)

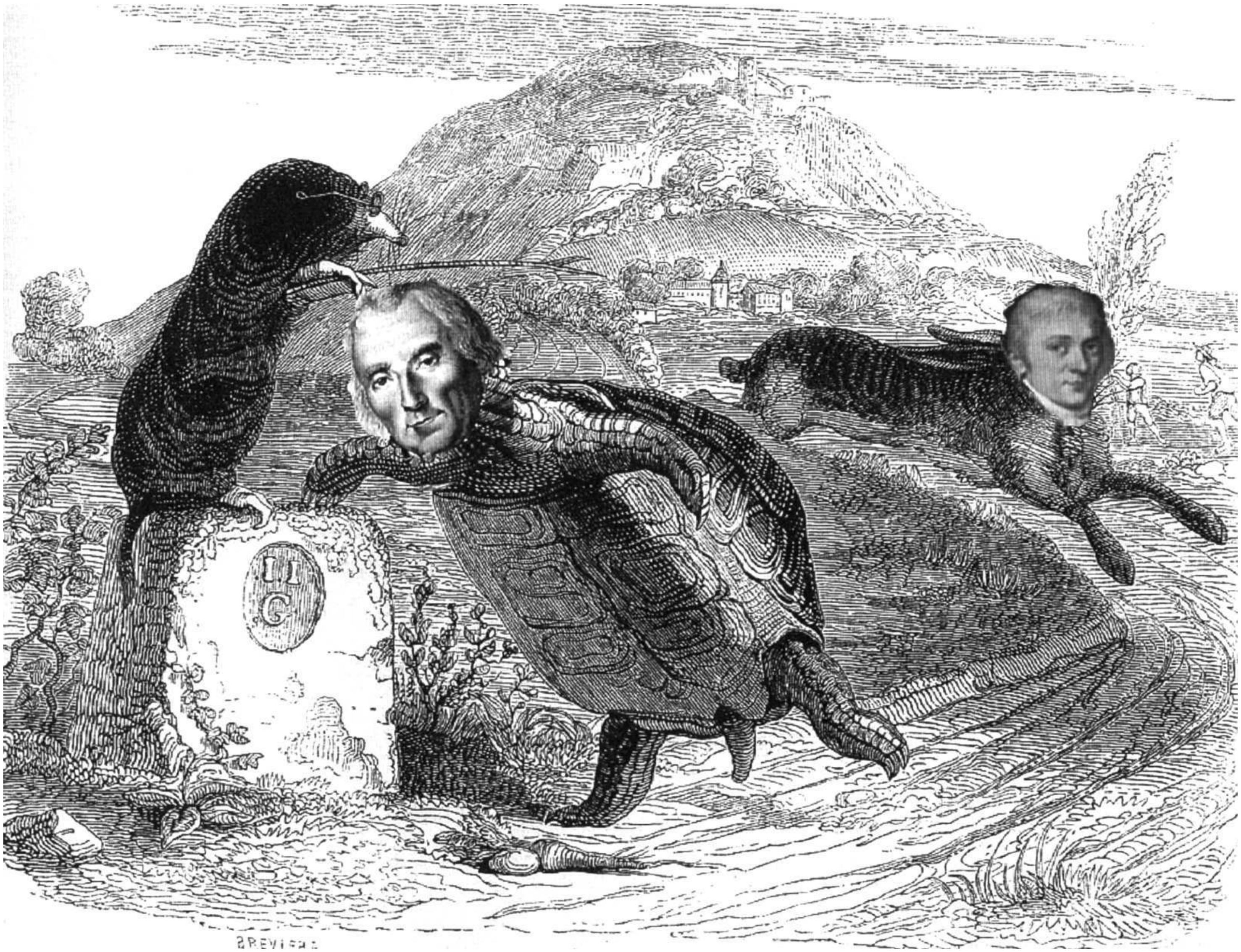
Regularizations

- **Main goal: avoid overfitting**

- **Two main lines of work**:

  1. Euclidean and Hilbertian norms (i.e., ℓ_2-norms)
     - Possibility of non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped theory and algorithms (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 and algorithms "in the making"**

ℓ_2 vs. ℓ_1 - Gaussian hare vs. Laplacian tortoise



- First-order methods (Fu, 1998; Beck and Teboulle, 2009)
- Homotopy methods (Markowitz, 1956; Efron et al., 2004)

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if there are low correlations between relevant and irrelevant variables.

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if there are low correlations between relevant and irrelevant variables.

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

$$\log p = O(n)$$

Going beyond the Lasso

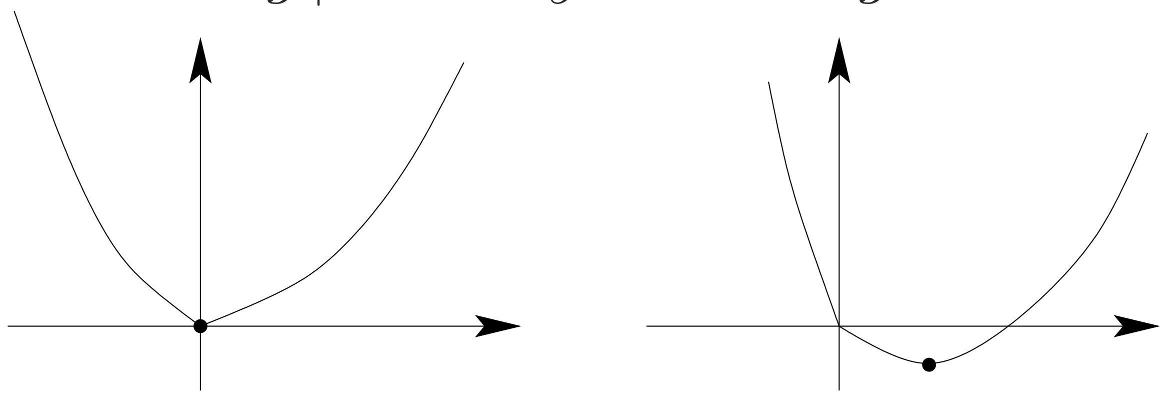
- ℓ_1-norm for **linear** feature selection in **high dimensions**
  - Lasso usually not applicable directly
- **Non-linearities**
- **Dealing with structured set of features**
- **Sparse learning on matrices**

Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Groups of features**
  - Non-linearity: Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)
- **Structured sparsity**
  - Overlapping groups and hierarchies

Why ℓ_1-norms lead to sparsity?

- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$

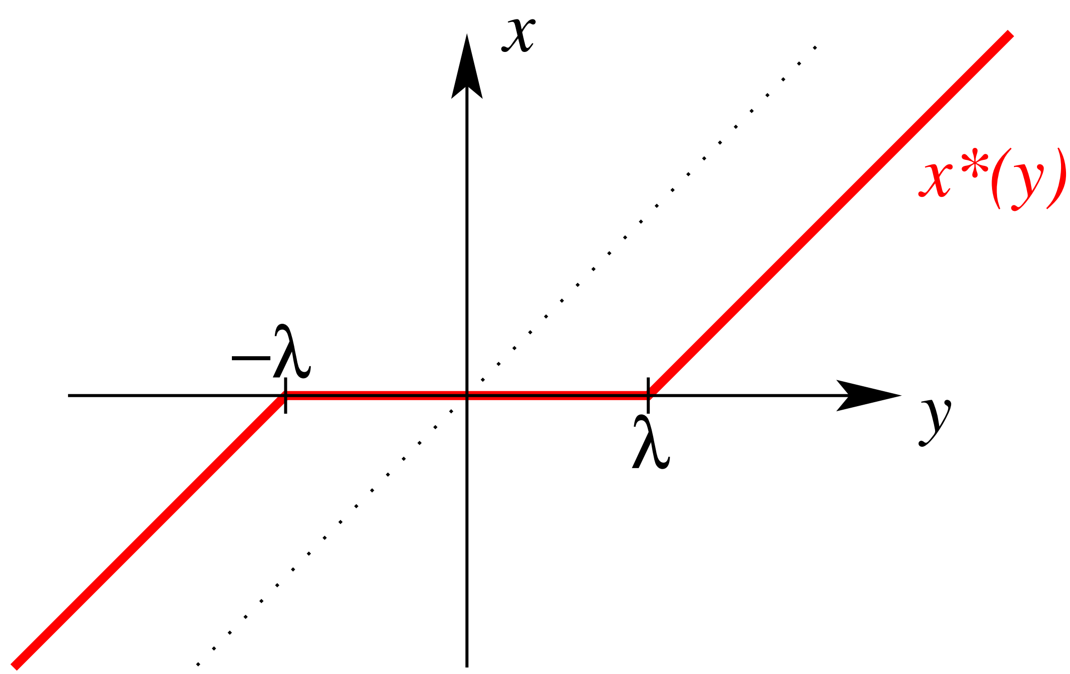


  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \leqslant 0$ (i.e., $y \leqslant -\lambda$) $\Rightarrow x^* = y + \lambda$

- Solution $\boxed{x^* = \operatorname{sign}(y)(|y| - \lambda)_+}$ = soft thresholding

Why ℓ_1-norms lead to sparsity?

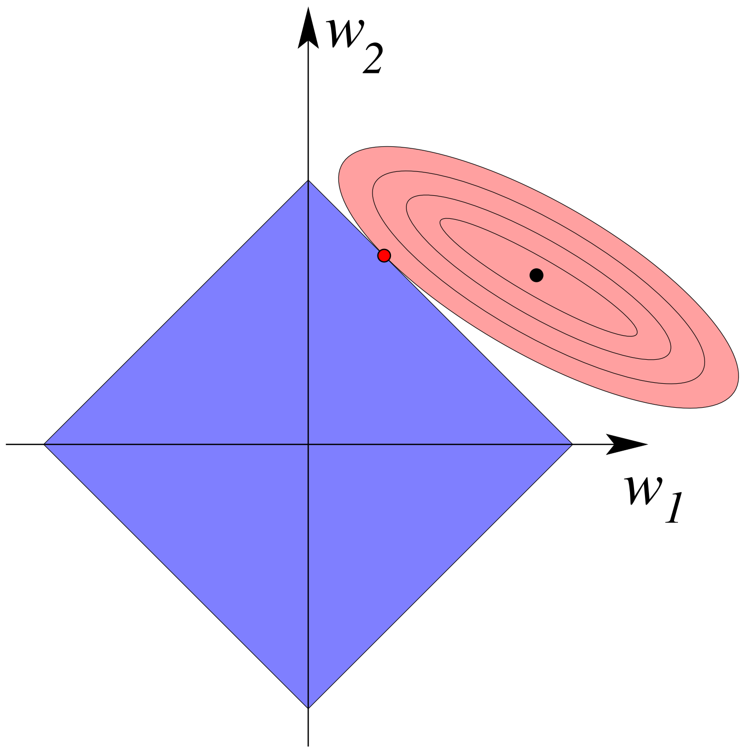
- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x\in\mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
- Solution $\boxed{x^* = \text{sign}(y)(|y| - \lambda)_+}$ = soft thresholding

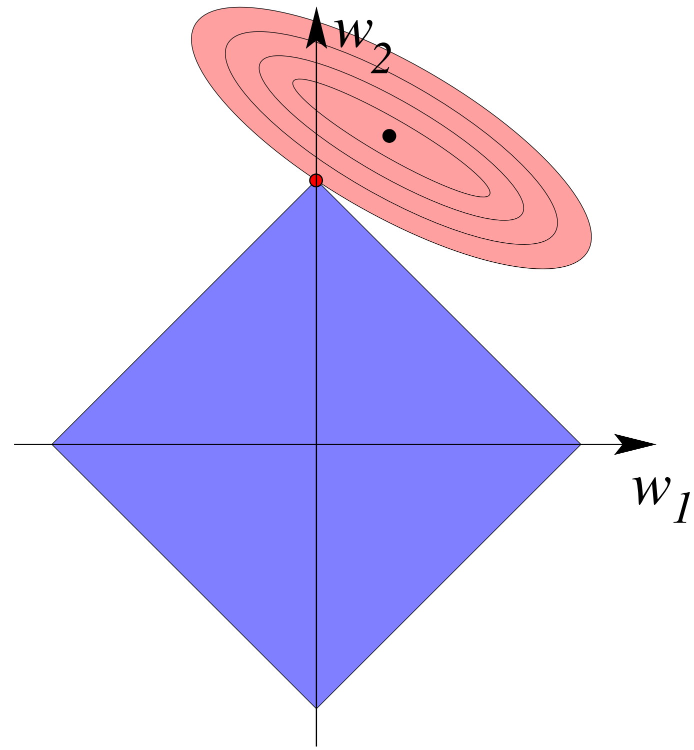


Why ℓ_1-norms lead to sparsity?

- **Example 2**: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding

- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining





ℓ_1-norm regularization (linear setting)

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

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

$$J(w) = \sum_{i=1}^{n} \ell(y_i, w^\top x_i) \quad + \quad \lambda \|w\|_{\mathbf{1}}$$

Error on data + Regularization

- Including a constant term b? Penalizing or constraining?

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

A review of nonsmooth convex analysis and optimization

- Analysis: optimality conditions
- Optimization: algorithms
  - First-order methods
- **Books**: Boyd and Vandenberghe (2004), Bonnans et al. (2003), Bertsekas (1995), Borwein and Lewis (2000)

Optimality conditions for smooth optimization
Zero gradient

- Example: ℓ_2-regularization: $\displaystyle \min_{w\in\mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top x_i) + \frac{\lambda}{2}\|w\|_2^2$
  - Gradient $\nabla J(w) = \sum_{i=1}^n \ell'(y_i, w^\top x_i)x_i + \lambda w$ where $\ell'(y_i, w^\top x_i)$ is the partial derivative of the loss w.r.t the second variable
  - If square loss, $\sum_{i=1}^n \ell(y_i, w^\top x_i) = \frac{1}{2}\|y - Xw\|_2^2$
    * gradient $= -X^\top(y - Xw) + \lambda w$
    * normal equations $\Rightarrow w = (X^\top X + \lambda I)^{-1}X^\top y$

Optimality conditions for smooth optimization

Zero gradient

- Example: ℓ_2-regularization: $\displaystyle \min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \frac{\lambda}{2} \|w\|_2^2$
  - Gradient $\nabla J(w) = \sum_{i=1}^n \ell'(y_i, w^\top x_i) x_i + \lambda w$ where $\ell'(y_i, w^\top x_i)$ is the partial derivative of the loss w.r.t the second variable
  - If square loss, $\sum_{i=1}^n \ell(y_i, w^\top x_i) = \frac{1}{2}\|y - Xw\|_2^2$
    * gradient $= -X^\top (y - Xw) + \lambda w$
    * normal equations $\Rightarrow w = (X^\top X + \lambda I)^{-1} X^\top y$

- **ℓ_1-norm is non differentiable!**
  - cannot compute the gradient of the absolute value

$\Rightarrow$ **Directional derivatives** (or subgradient)

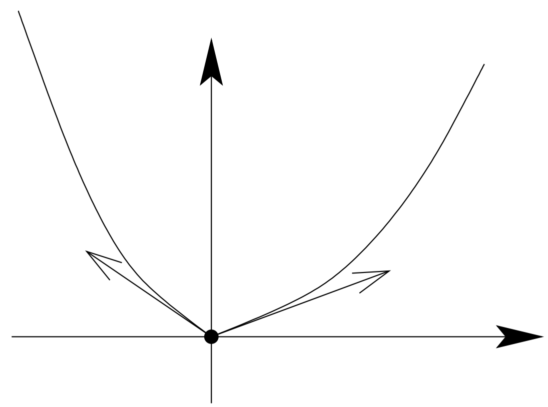
Directional derivatives - convex functions on $\mathbb{R}^p$

- Directional derivative in the direction Δ at w:

$$\nabla J(w, \Delta) = \lim_{\varepsilon \to 0+} \frac{J(w + \varepsilon\Delta) - J(w)}{\varepsilon}$$

- Always exist when J is convex and continuous

- Main idea: in non smooth situations, may need to look at all directions Δ and not simply p independent ones



- **Proposition**: J is differentiable at w, if and only if $\Delta \mapsto \nabla J(w, \Delta)$ is linear. Then, $\nabla J(w, \Delta) = \nabla J(w)^\top \Delta$

Optimality conditions for convex functions

- Unconstrained minimization (function defined on $\mathbb{R}^p$):
  - **Proposition**: w is optimal **if and only if** $\forall \Delta \in \mathbb{R}^p$, $\nabla J(w, \Delta) \geqslant 0$
  - Go up locally in all directions

- Reduces to zero-gradient for smooth problems

Directional derivatives for ℓ_1-norm regularization

- Function $J(w) = \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|_1 = L(w) + \lambda \|w\|_1$

- ℓ_1-norm: $\|w+\varepsilon\Delta\|_1 - \|w\|_1 = \sum_{j,\ w_j \neq 0} \{|w_j + \varepsilon\Delta_j| - |w_j|\} + \sum_{j,\ w_j = 0} |\varepsilon\Delta_j|$

- Thus,

$$\nabla J(w, \Delta) = \nabla L(w)^\top \Delta + \lambda \sum_{j,\ w_j \neq 0} \operatorname{sign}(w_j)\Delta_j + \lambda \sum_{j,\ w_j = 0} |\Delta_j|$$

$$= \sum_{j,\ w_j \neq 0} [\nabla L(w)_j + \lambda \operatorname{sign}(w_j)]\Delta_j + \sum_{j,\ w_j = 0} [\nabla L(w)_j \Delta_j + \lambda |\Delta_j|]$$

- Separability of optimality conditions

Optimality conditions for ℓ_1-norm regularization

- General loss: w optimal if and only if for all $j \in \{1, \dots, p\}$,

$$\begin{aligned} \operatorname{sign}(w_j) \neq 0 &\;\Rightarrow\; \nabla L(w)_j + \lambda \operatorname{sign}(w_j) = 0 \\ \operatorname{sign}(w_j) = 0 &\;\Rightarrow\; |\nabla L(w)_j| \leqslant \lambda \end{aligned}$$

- Square loss: w optimal if and only if for all $j \in \{1, \dots, p\}$,

$$\begin{aligned} \operatorname{sign}(w_j) \neq 0 &\;\Rightarrow\; -X_j^\top (y - Xw) + \lambda \operatorname{sign}(w_j) = 0 \\ \operatorname{sign}(w_j) = 0 &\;\Rightarrow\; |X_j^\top (y - Xw)| \leqslant \lambda \end{aligned}$$

  - For $J \subset \{1, \dots, p\}$, $X_J \in \mathbb{R}^{n \times |J|} = X(:, J)$ denotes the columns of X indexed by J, i.e., variables indexed by J

First order methods for convex optimization on $\mathbb{R}^p$

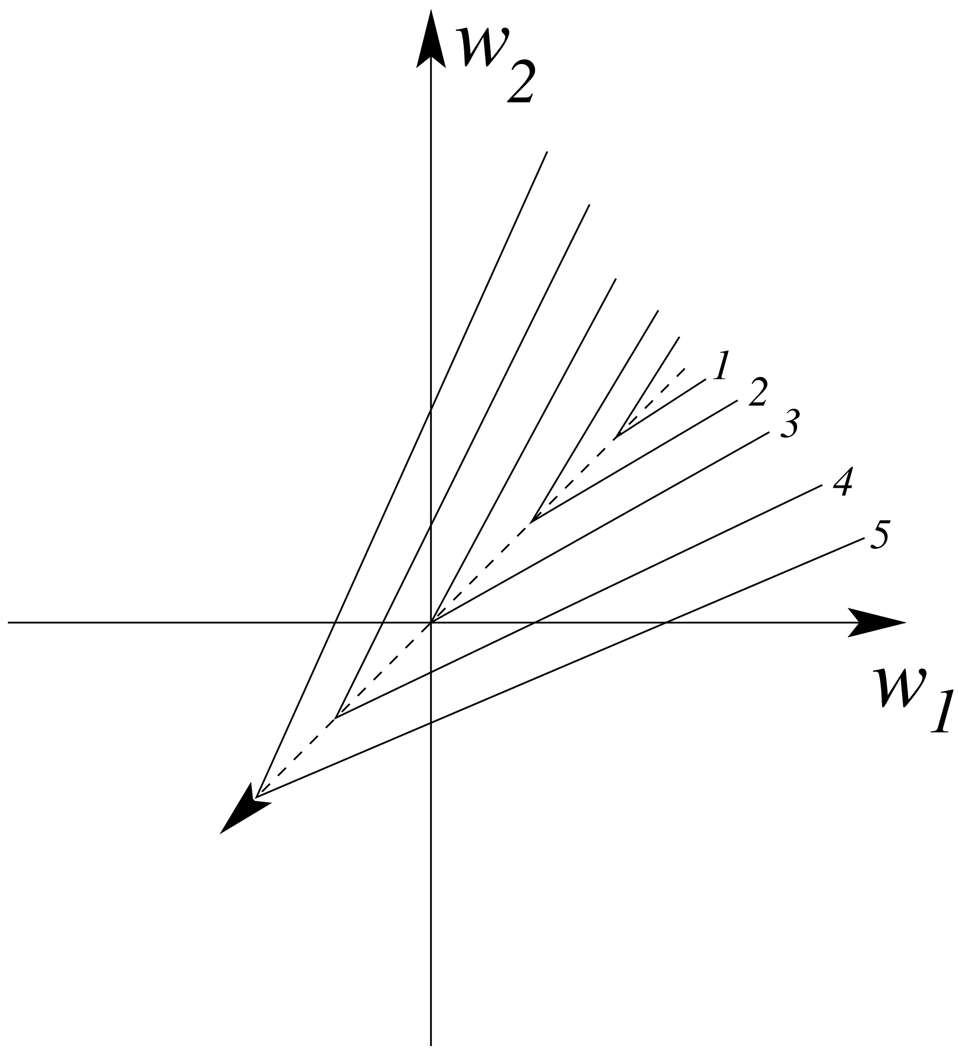
Smooth optimization

- **Gradient descent**: $w_{t+1} = w_t - \alpha_t \nabla J(w_t)$
  - with line search: search for a decent (not necessarily best) α_t
  - fixed diminishing step size, e.g., $\alpha_t = a(t+b)^{-1}$
- Convergence of $f(w_t)$ to $f^* = \min_{w \in \mathbb{R}^p} f(w)$ (Nesterov, 2003)
  - depends on condition number of the optimization problem (i.e., correlations within variables)
- **Coordinate descent**: similar properties

First order methods for convex optimization on $\mathbb{R}^p$
Smooth optimization

- **Gradient descent**: $w_{t+1} = w_t - \alpha_t \nabla J(w_t)$
  - with line search: search for a decent (not necessarily best) α_t
  - fixed diminishing step size, e.g., $\alpha_t = a(t+b)^{-1}$
- Convergence of $f(w_t)$ to $f^* = \min_{w \in \mathbb{R}^p} f(w)$ (Nesterov, 2003)
  - depends on condition number of the optimization problem (i.e., correlations within variables)
- **Coordinate descent**: similar properties
  - **Non-smooth objectives**: not always convergent

Counter-example
Coordinate descent for nonsmooth objectives



Regularized problems - Proximal methods

- Gradient descent as a proximal method (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{\mu}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{\mu}\nabla L(w_t)$

- Problems of the form: $\boxed{\min_{w \in \mathbb{R}^p} L(w) + \lambda\Omega(w)}$
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda\Omega(w) + \frac{\mu}{2}\|w - w_t\|_2^2$
  - Thresholded gradient descent $w_{t+1} = \text{SoftThres}(w_t - \frac{1}{\mu}\nabla L(w_t))$

- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)
  - **depends on the condition number of the loss**

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**
- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding

- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 2009)
  - Notice that $\sum_{j=1}^p |w_j| = \min_{\eta \geqslant 0} \frac{1}{2} \sum_{j=1}^p \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to η (closed-form $\eta_j = |w_j|$) and w (weighted squared ℓ_2-norm regularized problem)
  - Caveat: lack of continuity around $(w_i, \eta_i) = (0, 0)$: add ε/η_j

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding

- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 2009)
  - Notice that $\sum_{j=1}^{p} |w_j| = \min_{\eta \geqslant 0} \frac{1}{2} \sum_{j=1}^{p} \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to η (closed-form $\eta_j = |w_j|$) and w (weighted squared ℓ_2-norm regularized problem)
  - Caveat: lack of continuity around $(w_i, \eta_i) = (0, 0)$: add ε / η_i

- **Dedicated algorithms that use sparsity** (active sets/homotopy)

Special case of square loss

- **Quadratic programming formulation**: minimize

$$\frac{1}{2}\|y - Xw\|^2 + \lambda \sum_{j=1}^{p} (w_j^+ + w_j^-) \text{ s.t. } w = w^+ - w^-, \; w^+ \geqslant 0, \; w^- \geqslant 0$$

Special case of square loss

- **Quadratic programming formulation**: minimize

$$\frac{1}{2}\|y - Xw\|^2 + \lambda \sum_{j=1}^{p} (w_j^+ + w_j^-) \text{ s.t. } w = w^+ - w^-,\ w^+ \geqslant 0,\ w^- \geqslant 0$$

  - **generic toolboxes ⇒ very slow**

- **Main property**: if the sign pattern $s \in \{-1, 0, 1\}^p$ of the solution is known, the solution can be obtained in closed form
  - Lasso equivalent to minimizing $\frac{1}{2}\|y - X_J w_J\|^2 + \lambda s_J^\top w_J$ w.r.t. w_J where $J = \{j, s_j \neq 0\}$.
  - Closed form solution $w_J = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$

- **Algorithm: "Guess" s and check optimality conditions**

Optimality conditions for ℓ_1-norm regularization

- General loss: w optimal if and only if for all $j \in \{1, \dots, p\}$,

$$\begin{aligned} \operatorname{sign}(w_j) \neq 0 \;&\Rightarrow\; \nabla L(w)_j + \lambda \operatorname{sign}(w_j) = 0 \\ \operatorname{sign}(w_j) = 0 \;&\Rightarrow\; |\nabla L(w)_j| \leqslant \lambda \end{aligned}$$

- Square loss: w optimal if and only if for all $j \in \{1, \dots, p\}$,

$$\begin{aligned} \operatorname{sign}(w_j) \neq 0 \;&\Rightarrow\; -X_j^\top (y - Xw) + \lambda \operatorname{sign}(w_j) = 0 \\ \operatorname{sign}(w_j) = 0 \;&\Rightarrow\; |X_j^\top (y - Xw)| \leqslant \lambda \end{aligned}$$

  - For $J \subset \{1, \dots, p\}$, $X_J \in \mathbb{R}^{n \times |J|} = X(:, J)$ denotes the columns of X indexed by J, i.e., variables indexed by J

Optimality conditions for the sign vector s (Lasso)

- For $s \in \{-1, 0, 1\}^p$ sign vector, $J = \{j, s_j \neq 0\}$ the nonzero pattern
- potential closed form solution: $w_J = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$ and $w_{J^c} = 0$
- s is optimal if and only if
  - active variables: $\text{sign}(w_J) = s_J$
  - inactive variables: $\|X_{J^c}^\top (y - X_J w_J)\|_\infty \leqslant \lambda$
- **Active set algorithms** (Lee et al., 2007; Roth and Fischer, 2008)
  - Construct J iteratively by adding variables to the active set
  - Only requires to invert small linear systems

Homotopy methods for the square loss (Markowitz, 1956; Osborne et al., 2000; Efron et al., 2004)

- **Goal**: Get all solutions for all possible values of the regularization parameter λ

- Same idea as before: if the sign vector is known,

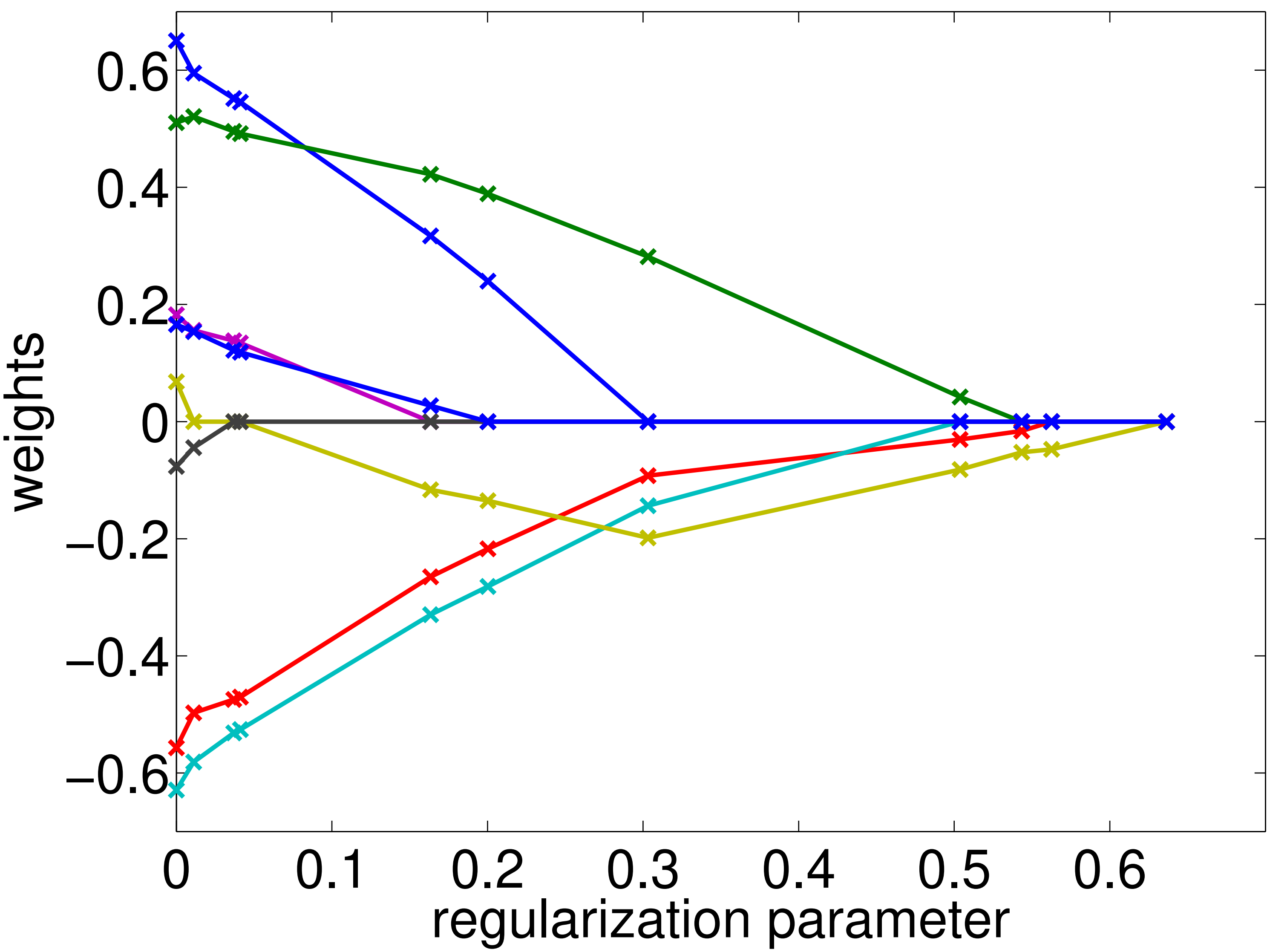
$$w_J^*(\lambda) = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$$

  valid, as long as,

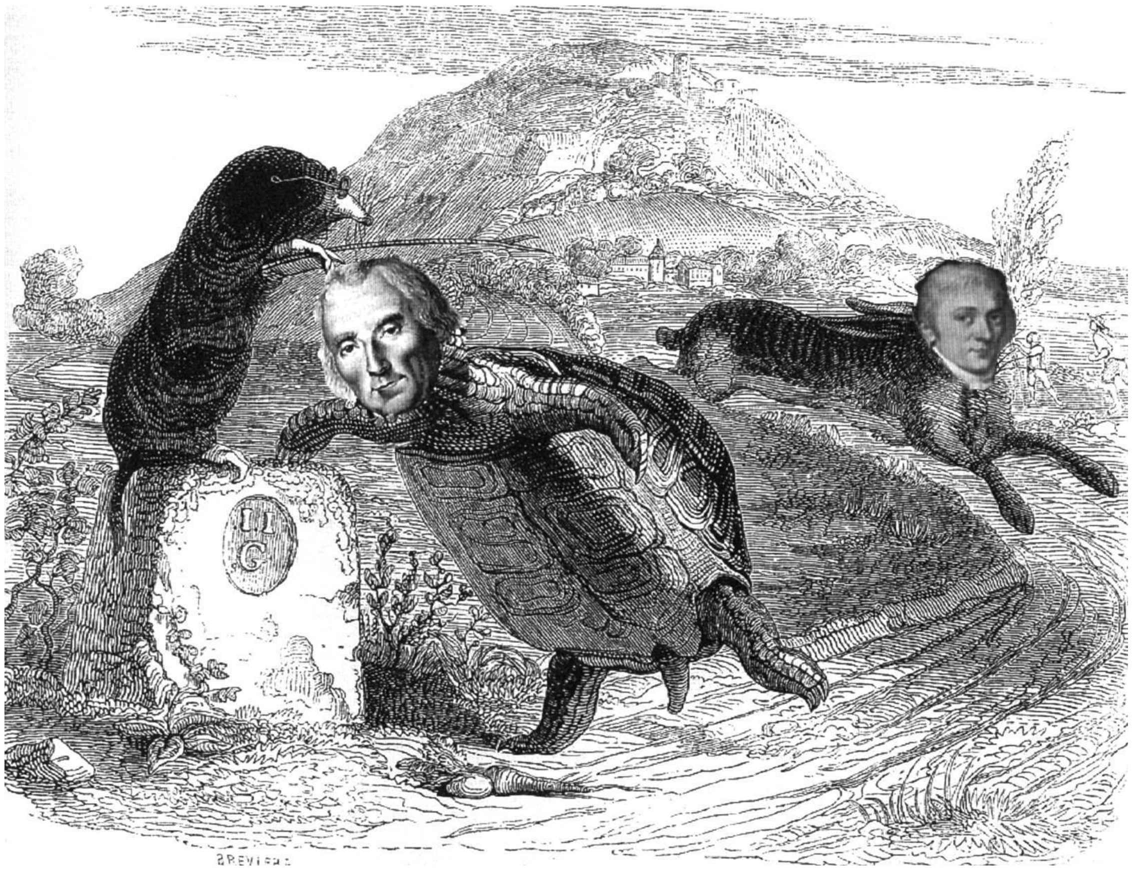
  - sign condition: $\text{sign}(w_J^*(\lambda)) = s_J$
  - subgradient condition: $\|X_{J^c}^\top (X_J w_J^*(\lambda) - y)\|_\infty \leqslant \lambda$
  - this defines an interval on λ: the path is thus **piecewise affine**

- Simply need to find break points and directions

Piecewise linear paths



Algorithms for ℓ_1-norms (square loss): Gaussian hare vs. Laplacian tortoise



- Coord. descent and proximal: $O(pn)$ per iterations for ℓ_1 and ℓ_2
- "Exact" algorithms: $O(kpn)$ for ℓ_1 **vs.** $O(p^2n)$ for ℓ_2

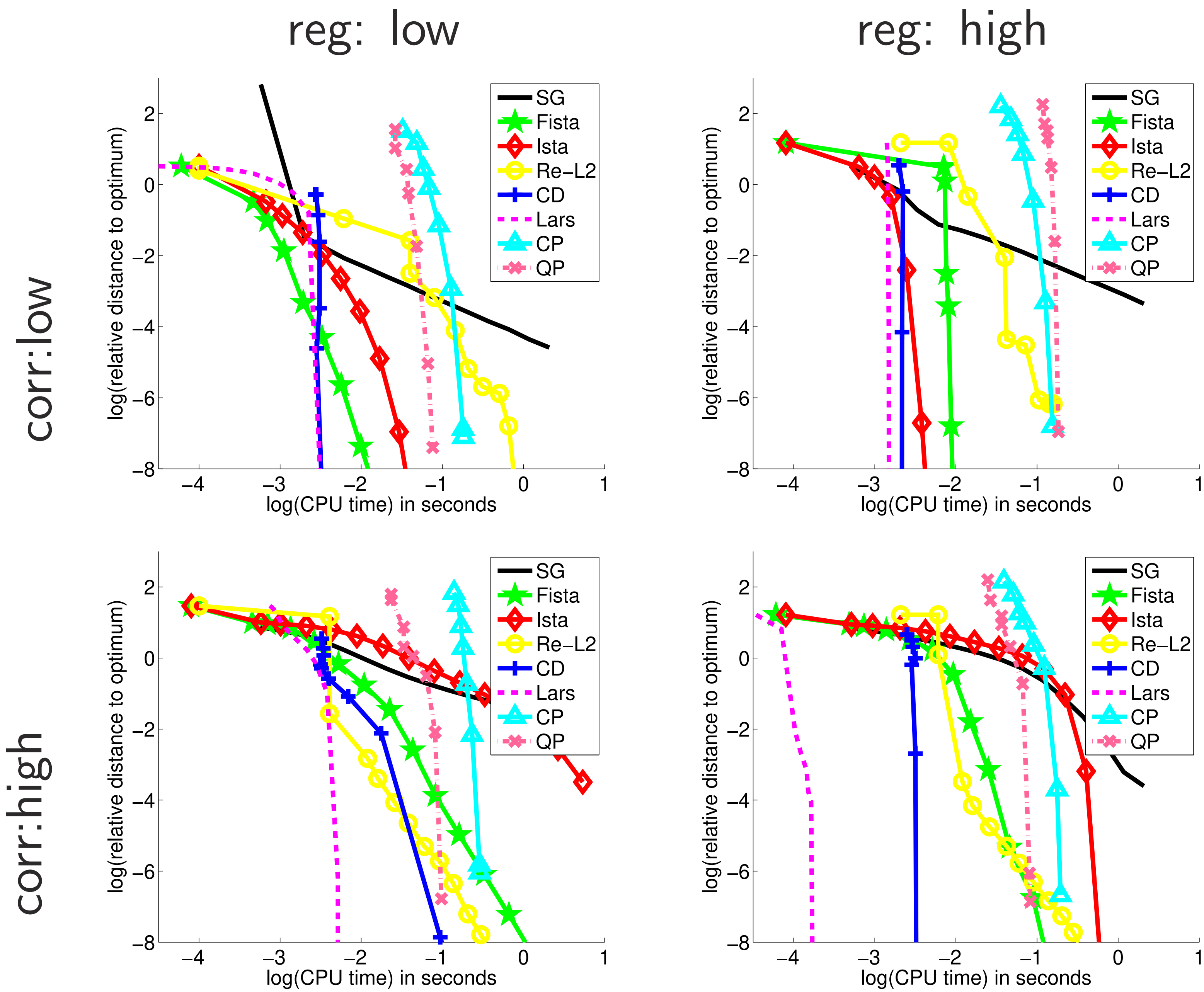
Additional methods - Softwares

- Many contributions in signal processing, optimization, mach. learning
  - Extensions to stochastic setting (Bottou and Bousquet, 2008)

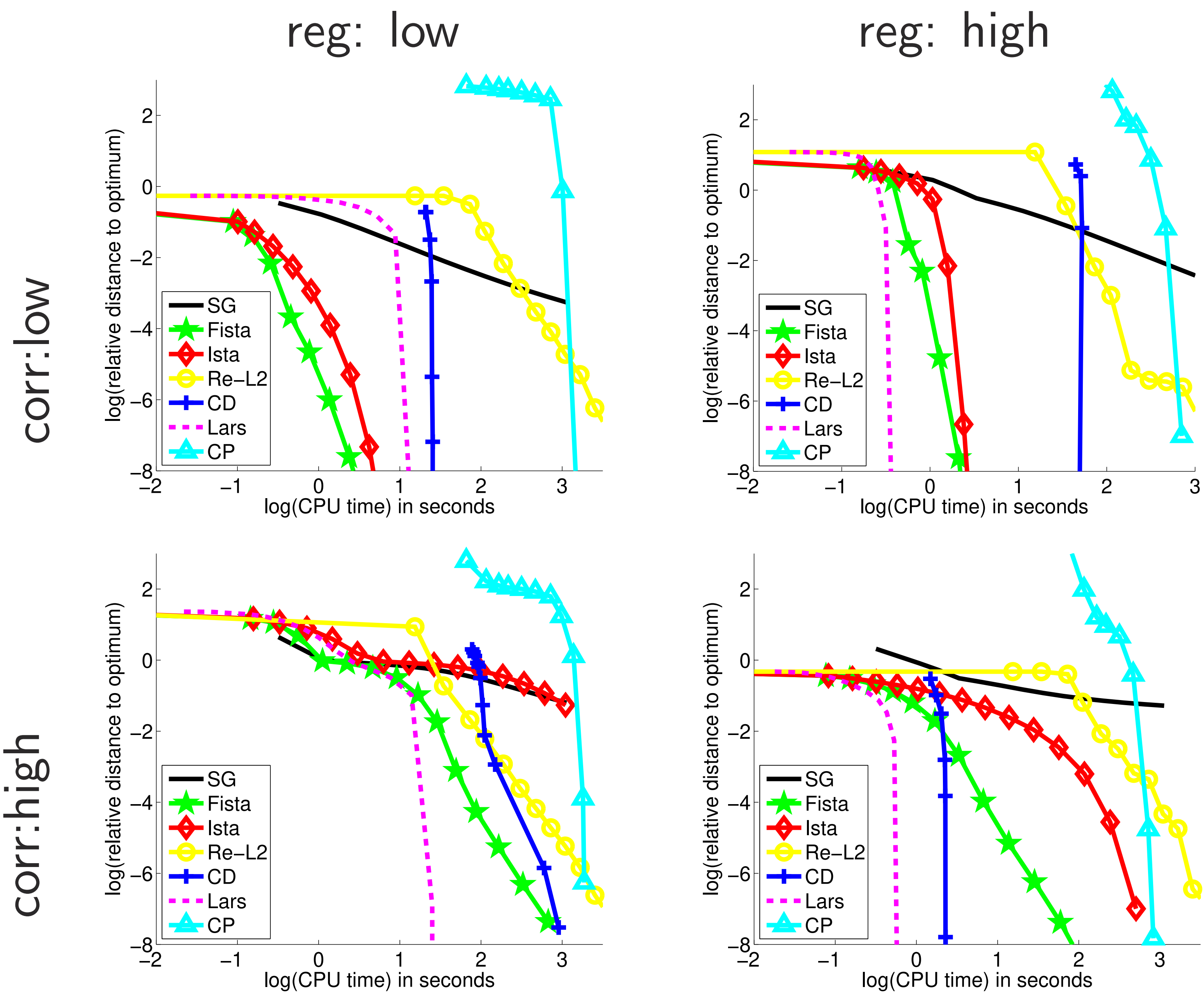
- **Extensions to other sparsity-inducing norms**
  - Computing proximal operator
  - F. Bach, R. Jenatton, J. Mairal, G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1-106, 2011.

- **Softwares**
  - Many available codes
  - SPAMS (SPArse Modeling Software)
    http://www.di.ens.fr/willow/SPAMS/

Empirical comparison: small scale ($n = 200$, $p = 200$)



Empirical comparison: medium scale ($n = 2000$, $p = 10000$)



Empirical comparison: conclusions

- **Lasso**
  - Generic methods very slow
  - LARS fastest in **low dimension** or for **high correlation**
  - Proximal methods competitive
    - especially larger setting with weak corr. + weak reg.
  - Coordinate descent
    - Dominated by the LARS
    - Would benefit from an offline computation of the matrix

- **Smooth Losses**
  - LARS not available $\rightarrow$ CD and proximal methods good candidates

Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Groups of features**
  - Non-linearity: Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)
- **Structured sparsity**
  - Overlapping groups and hierarchies

Theoretical results - Square loss

- Main assumption: data generated from a certain sparse $\mathbf{w}$

- Three main problems:

  1. **Regular consistency**: convergence of estimator $\hat{w}$ to $\mathbf{w}$, i.e., $\|\hat{w} - \mathbf{w}\|$ tends to zero when n tends to ∞
  2. **Model selection consistency**: convergence of the sparsity pattern of $\hat{w}$ to the pattern $\mathbf{w}$
  3. **Efficiency**: convergence of predictions with $\hat{w}$ to the predictions with $\mathbf{w}$, i.e., $\frac{1}{n}\|X\hat{w} - X\mathbf{w}\|_2^2$ tends to zero

- Main results:

  – **Condition for model consistency (support recovery)**
  – **High-dimensional inference**

Model selection consistency (Lasso)

- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern
- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

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

where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

Model selection consistency (Lasso)

- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern
- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

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

  where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \text{Supp}(\mathbf{w})$
- Condition depends on $\mathbf{w}$ and $\mathbf{J}$ (may be relaxed)
  - may be relaxed by maximizing out $\text{sign}(\mathbf{w})$ or $\mathbf{J}$
- Valid in low and high-dimensional settings
- Requires lower-bound on magnitude of nonzero $\mathbf{w}_j$

Model selection consistency (Lasso)

- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern

- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

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

  where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

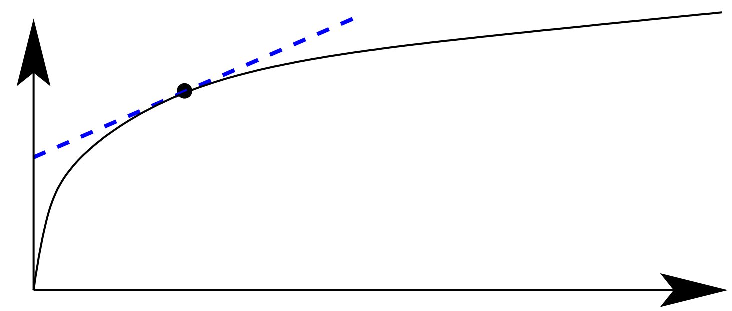
- **The Lasso is usually not model-consistent**
  - Selects more variables than necessary (see, e.g., Lv and Fan, 2009)
  - **Fixing the Lasso**: adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2008), thresholding (Lounici, 2008), Bolasso (Bach, 2008a), stability selection (Meinshausen and Bühlmann, 2008), Wasserman and Roeder (2009)

Adaptive Lasso and concave penalization

- **Adaptive Lasso** (Zou, 2006; Huang et al., 2008)
  - Weighted ℓ_1-norm: $\min_{w \in \mathbb{R}^p} L(w) + \lambda \sum_{j=1}^{p} \frac{|w_j|}{|\hat{w}_j|^\alpha}$
  - $\hat{w}$ estimator obtained from ℓ_2 or ℓ_1 regularization

- **Reformulation in terms of concave penalization**

$$\min_{w \in \mathbb{R}^p} L(w) + \sum_{j=1}^{p} g(|w_j|)$$

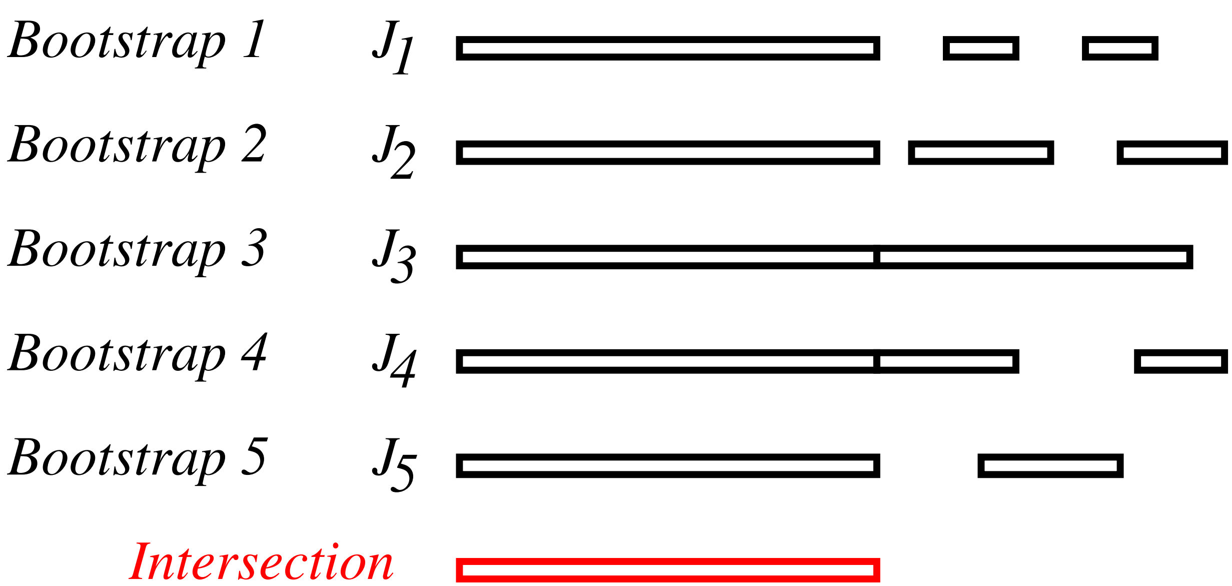


  - Example: $g(|w_j|) = |w_j|^{1/2}$ or $\log |w_j|$. Closer to the ℓ_0 penalty
  - Concave-convex procedure: replace $g(|w_j|)$ by affine upper bound
  - Better sparsity-inducing properties (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2008b)

Bolasso (Bach, 2008a)

- **Property**: for a specific choice of regularization parameter $\lambda \approx \sqrt{n}$:
  - all variables in $\mathbf{J}$ are always selected with high probability
  - all other ones selected with probability in $(0, 1)$

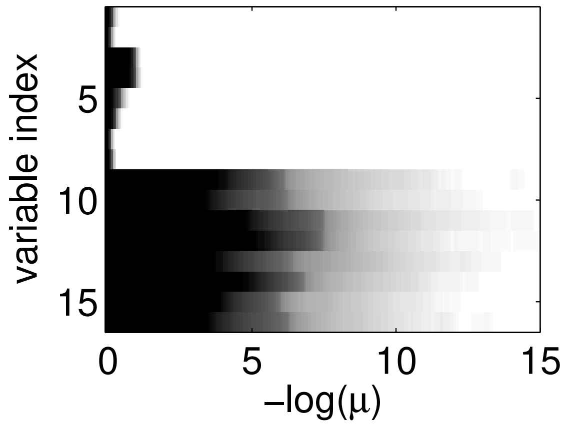
- Use the bootstrap to simulate several replications
  - Intersecting supports of variables
  - Final estimation of w on the entire dataset

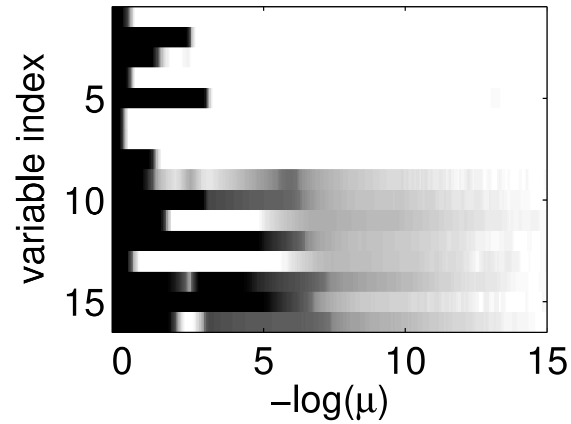


Model selection consistency of the Lasso/Bolasso

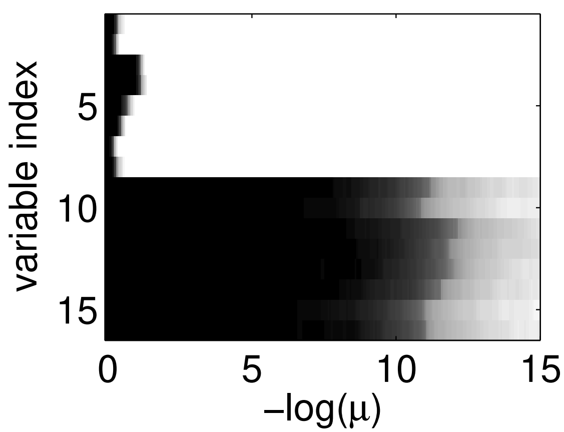
- probabilities of selection of each variable vs. regularization param. μ

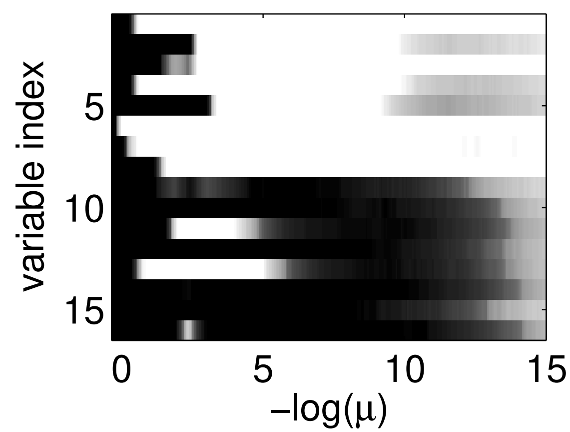
LASSO





BOLASSO





Support recovery condition satisfied not satisfied

High-dimensional inference
Going beyond exact support recovery

- Theoretical results usually assume that non-zero $\mathbf{w}_j$ are large enough, i.e., $|\mathbf{w}_j| \geqslant \sigma\sqrt{\frac{\log p}{n}}$
- **May include too many variables but still predict well**
- Oracle inequalities
  - Predict as well as the estimator obtained with the knowledge of $\mathbf{J}$
  - Assume i.i.d. Gaussian noise with variance σ^2
  - We have:
$$\frac{1}{n}\mathbb{E}\|X\hat{w}_{\text{oracle}} - X\mathbf{w}\|_2^2 = \frac{\sigma^2|J|}{n}$$

High-dimensional inference
Variable selection without computational limits

- Approaches based on penalized criteria (close to BIC)

$$\min_{w \in \mathbb{R}^p} \tfrac{1}{2}\|y - Xw\|_2^2 + C\sigma^2 \|w\|_0 \big(1 + \log \frac{p}{\|w\|_0}\big)$$

- **Oracle inequality** if data generated by $\mathbf{w}$ with k non-zeros (Massart, 2003; Bunea et al., 2007):

$$\frac{1}{n}\|X\hat{w} - X\mathbf{w}\|_2^2 \leqslant C \frac{k\sigma^2}{n} \big(1 + \log \frac{p}{k}\big)$$

- Gaussian noise - **No assumptions regarding correlations**

- **Scaling between dimensions**: $\dfrac{k \log p}{n}$ small

High-dimensional inference (Lasso)

- **Main result**: we only need $k \log p = O(n)$
  - if $\mathbf{w}$ is sufficiently sparse
  - **and** input variables are not too correlated

High-dimensional inference (Lasso)

- **Main result**: we only need $k \log p = O(n)$
  - if $\mathbf{w}$ is sufficiently sparse
  - **and** input variables are not too correlated
- Precise conditions on covariance matrix $\mathbf{Q} = \frac{1}{n} X^\top X$.
  - **Mutual incoherence** (Lounici, 2008)
  - Restricted eigenvalue conditions (Bickel et al., 2009)
  - Sparse eigenvalues (Meinshausen and Yu, 2008)
  - Null space property (Donoho and Tanner, 2005)
- Links with signal processing and compressed sensing (Candès and Wakin, 2008)

Mutual incoherence (uniform low correlations)

- **Theorem** (Lounici, 2008):
  - $y_i = \mathbf{w}^\top x_i + \varepsilon_i$, ε i.i.d. normal with mean zero and variance σ^2
  - $\mathbf{Q} = X^\top X/n$ with unit diagonal and cross-terms less than $\frac{1}{14k}$
  - if $\|\mathbf{w}\|_0 \leqslant k$, and $A^2 > 8$, then, with $\lambda = A\sigma\sqrt{n \log p}$

$$\mathbb{P}\left(\|\hat{w} - \mathbf{w}\|_\infty \leqslant 5A\sigma \left(\frac{\log p}{n} \right)^{1/2} \right) \geqslant 1 - p^{1-A^2/8}$$

- Model consistency by thresholding if $\min_{j, \mathbf{w}_j \neq 0} |\mathbf{w}_j| > C\sigma\sqrt{\frac{\log p}{n}}$
- Mutual incoherence condition depends *strongly* on k
- Improved result by averaging over sparsity patterns (Candès and Plan, 2009)

Restricted eigenvalue conditions

- **Theorem** (Bickel et al., 2009):
  - assume $\boxed{\kappa(k)^2 = \min_{|J|\leqslant k} \min_{\Delta,\ \|\Delta_{J^c}\|_1 \leqslant \|\Delta_J\|_1} \frac{\Delta^\top \mathbf{Q} \Delta}{\|\Delta_J\|_2^2} > 0}$
  - assume $\lambda = A\sigma\sqrt{n \log p}$ and $A^2 > 8$
  - then, with probability $1 - p^{1-A^2/8}$, we have

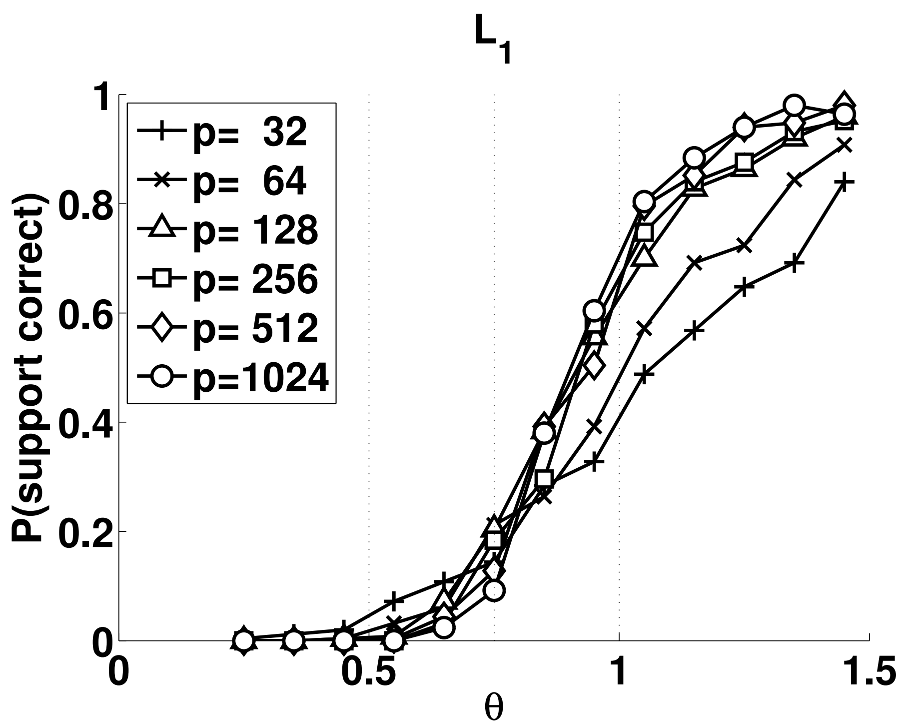
$$\text{estimation error} \qquad \|\hat{w} - \mathbf{w}\|_1 \leqslant \frac{16A}{\kappa^2(k)} \sigma k \sqrt{\frac{\log p}{n}}$$

$$\text{prediction error} \qquad \frac{1}{n}\|X\hat{w} - X\mathbf{w}\|_2^2 \leqslant \frac{16A^2}{\kappa^2(k)} \frac{\sigma^2 k}{n} \log p$$

- Condition imposes a potentially hidden scaling between (n, p, k)
- Condition always satisfied for $\mathbf{Q} = I$

Checking sufficient conditions

- **Most of the conditions are not computable in polynomial time**
- **Random matrices**
  - Sample $X \in \mathbb{R}^{n \times p}$ from the Gaussian ensemble
  - Conditions satisfied with high probability for certain (n, p, k)
  - Example from Wainwright (2009): $\boxed{\theta = \frac{n}{2k \log p} > 1}$



Sparse methods
Common extensions

- **Removing bias of the estimator**
  - Keep the active set, and perform unregularized restricted estimation (Candès and Tao, 2007)
  - Better theoretical bounds
  - Potential problems of robustness
- **Elastic net** (Zou and Hastie, 2005)
  - Replace $\lambda\|w\|_1$ by $\lambda\|w\|_1 + \varepsilon\|w\|_2^2$
  - Make the optimization strongly convex with unique solution
  - Better behavior with heavily correlated variables

Relevance of theoretical results

- **Most results only for the square loss**
  - Extend to other losses (Van De Geer, 2008; Bach, 2009)
- **Most results only for ℓ_1-regularization**
  - May be extended to other norms (see, e.g., Huang and Zhang, 2009; Bach, 2008b)
- **Condition on correlations**
  - very restrictive, far from results for BIC penalty
- **Non sparse generating vector**
  - little work on robustness to lack of sparsity
- **Estimation of regularization parameter**
  - No satisfactory solution $\Rightarrow$ open problem

Alternative sparse methods
Greedy methods

- Forward selection
- Forward-backward selection
- Non-convex method
  - Harder to analyze
  - Simpler to implement
  - Problems of stability
- Positive theoretical results (Zhang, 2009, 2008a)
  - Similar sufficient conditions than for the Lasso

Alternative sparse methods
Bayesian methods

- Lasso: minimize $\sum_{i=1}^n (y_i - w^\top x_i)^2 + \lambda \|w\|_1$
  - Equivalent to MAP estimation with Gaussian likelihood and factorized **Laplace** prior $p(w) \propto \prod_{j=1}^p e^{-\lambda |w_j|}$ (Seeger, 2008)
  - **However, posterior puts zero weight on exact zeros**

- Heavy-tailed distributions as a proxy to sparsity
  - Student distributions (Caron and Doucet, 2008)
  - Generalized hyperbolic priors (Archambeau and Bach, 2008)
  - Instance of automatic relevance determination (Neal, 1996)

- Mixtures of "Diracs" and another absolutely continuous distributions, e.g., "spike and slab" (Ishwaran and Rao, 2005)

- Less theory than frequentist methods

Comparing Lasso and other strategies for linear regression

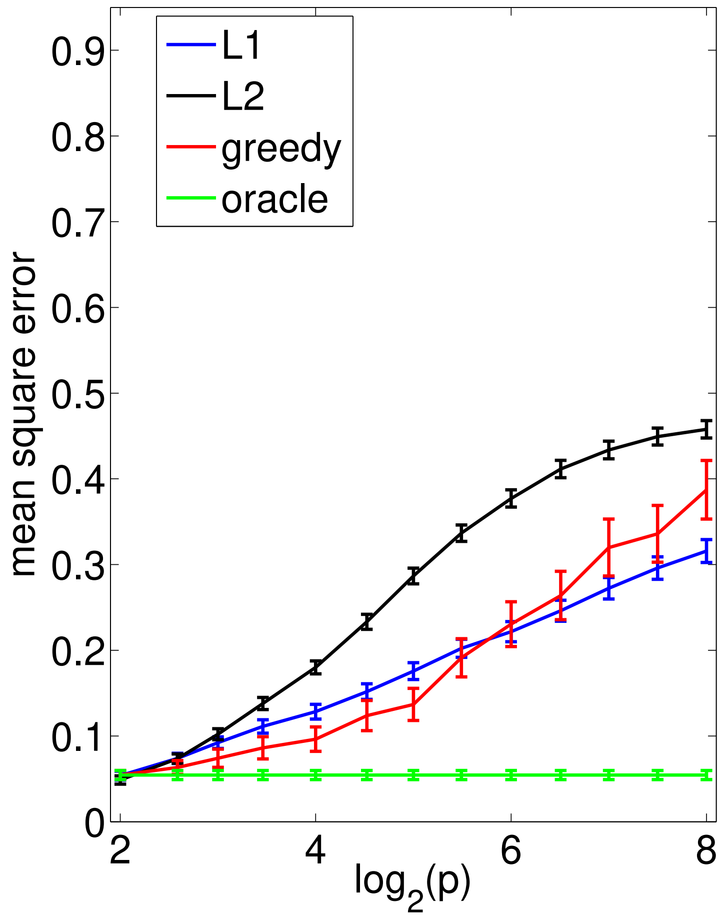
- Compared methods to reach the least-square solution
  - Ridge regression: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
  - Lasso: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$
  - Forward greedy:
    - Initialization with empty set
    - Sequentially add the variable that best reduces the square loss

- Each method builds a path of solutions from 0 to ordinary least-squares solution

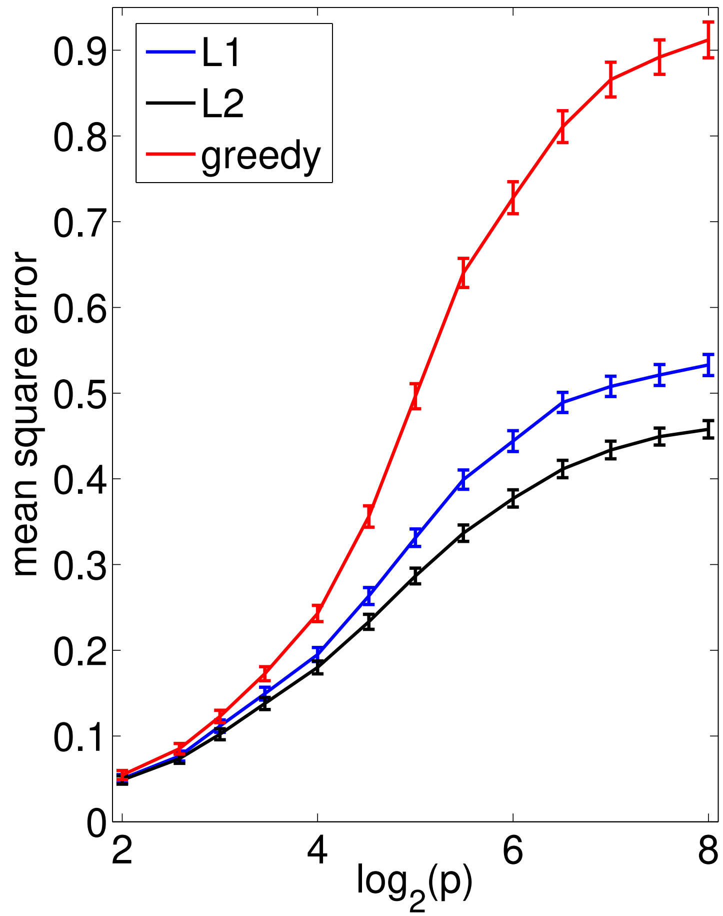
- Regularization parameters selected on the test set

Simulation results

- i.i.d. Gaussian design matrix, $k = 4$, $n = 64$, $p \in [2, 256]$, SNR $= 1$
- Note stability to non-sparsity and variability



Sparse



Rotated (non sparse)

Summary

ℓ_1-norm regularization

- ℓ_1-norm regularization leads to **nonsmooth optimization problems**
  - analysis through directional derivatives or subgradients
  - optimization may or may not take advantage of sparsity
- ℓ_1-norm regularization allows **high-dimensional inference**
- Interesting problems for ℓ_1-regularization
  - Stable variable selection
  - Weaker sufficient conditions (for weaker results)
  - Estimation of regularization parameter (all bounds depend on the unknown noise variance σ^2)

Extensions

- **Sparse methods are not limited to the square loss**
  - logistic loss: algorithms (Beck and Teboulle, 2009) and theory (Van De Geer, 2008; Bach, 2009)

- **Sparse methods are not limited to supervised learning**
  - Learning the structure of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008)
  - Sparsity on matrices (last part of the tutorial)

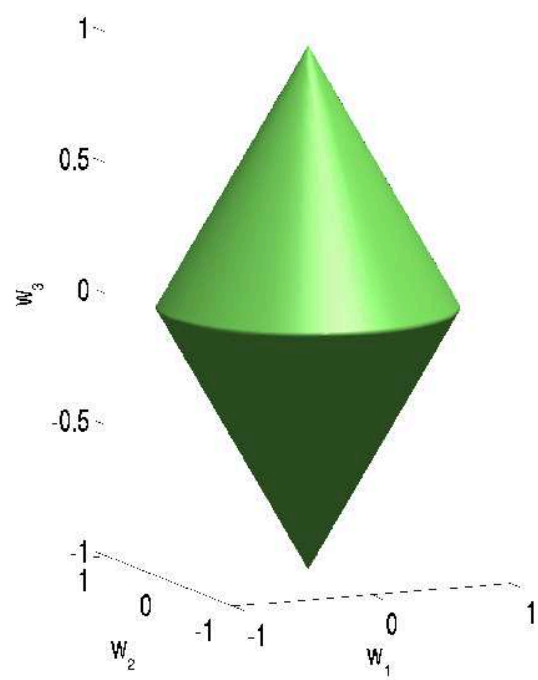
- **Sparse methods are not limited to variable selection in a linear model**
  - **See next parts of the tutorial**

Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Groups of features**
  - Non-linearity: Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)
- **Structured sparsity**
  - Overlapping groups and hierarchies

Penalization with grouped variables (Yuan and Lin, 2006)

- Assume that $\{1, \ldots, p\}$ is **partitioned** into m groups $G_1, \ldots, G_m$
- Penalization by $\sum_{i=1}^{m} \|w_{G_i}\|_2$, often called ℓ_1-ℓ_2 norm
- Induces group sparsity
  - Some groups entirely set to zero
  - no zeros within groups
  - Unit ball in $\mathbb{R}^3$: $\|(w_1, w_2)\| + \|w_3\| \leq 1$
- In this tutorial:
  - Groups may have infinite size $\Rightarrow$ **MKL**
  - Groups may overlap $\Rightarrow$ **structured sparsity**



Linear vs. non-linear methods

- All methods in this tutorial are **linear in the parameters**
- By replacing x by features $\Phi(x)$, they can be made **non linear in the data**
- **Implicit vs. explicit features**
  - ℓ_1-norm: explicit features
  - ℓ_2-norm: representer theorem allows to consider implicit features if their dot products can be computed easily (kernel methods)

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

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:

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

- **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:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha$$

  - 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 of 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 of 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

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)

General kernel learning

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

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

  is a **convex** function of the kernel matrix K

- Theoretical learning bounds (Lanckriet et al., 2004, Srebro and Ben-David, 2006)

General kernel learning

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

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

  is a **convex** function of the kernel matrix K

- Theoretical learning bounds (Lanckriet et al., 2004, Srebro and Ben-David, 2006)

- Natural parameterization $\boxed{K = \sum_{j=1}^{p} \eta_j K_j}$, $\eta \geqslant 0$, $\sum_{j=1}^{p} \eta_j = 1$
  - Interpretation in terms of group sparsity

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 $x \in \mathbb{R}^p$ and $w_j \in \mathbb{R}$

  – by $f(x) = \sum_{j=1}^{p} w_j^\top \Phi_j(x)$ with $x \in \mathcal{X}$, $\Phi_j(x) \in \mathcal{F}_j$ 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)

Regularization for multiple features

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

- 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

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

- 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
  2. Analysis of sparsity inducing properties (Ravikumar et al., 2008; Bach, 2008b)
  3. Does it correspond to a specific combination of kernels?

Equivalence with kernel learning (Bach et al., 2004a)

- Block ℓ_1-norm problem:

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

- **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**

Proof of equivalence

$$\min_{w_1,\dots,w_p} \sum_{i=1}^{n} \ell\Big(y_i, \sum_{j=1}^{p} w_j^\top \Phi_j(x_i)\Big) + \lambda\Big(\sum_{j=1}^{p} \|w_j\|_2\Big)^2$$

$$= \min_{w_1,\dots,w_p} \min_{\sum_j \eta_j = 1} \sum_{i=1}^{n} \ell\Big(y_i, \sum_{j=1}^{p} w_j^\top \Phi_j(x_i)\Big) + \lambda \sum_{j=1}^{p} \|w_j\|_2^2/\eta_j$$

$$= \min_{\sum_j \eta_j = 1} \min_{\tilde{w}_1,\dots,\tilde{w}_p} \sum_{i=1}^{n} \ell\Big(y_i, \sum_{j=1}^{p} \eta_j^{1/2} \tilde{w}_j^\top \Phi_j(x_i)\Big) + \lambda \sum_{j=1}^{p} \|\tilde{w}_j\|_2^2 \text{ with } \tilde{w}_j = w_j \eta_j^{-1/2}$$

$$= \min_{\sum_j \eta_j = 1} \min_{\tilde{w}} \sum_{i=1}^{n} \ell\big(y_i, \tilde{w}^\top \Psi_\eta(x_i)\big) + \lambda \|\tilde{w}\|_2^2 \text{ with } \Psi_\eta(x) = (\eta_1^{1/2}\Phi_1(x), \dots, \eta_p^{1/2}\Phi_p(x))$$

- We have: $\Psi_\eta(x)^\top \Psi_\eta(x') = \sum_{j=1}^{p} \eta_j k_j(x, x')$ with $\sum_{j=1}^{p} \eta_j = 1$ (and $\eta \geqslant 0$)

Algorithms for the group Lasso / MKL

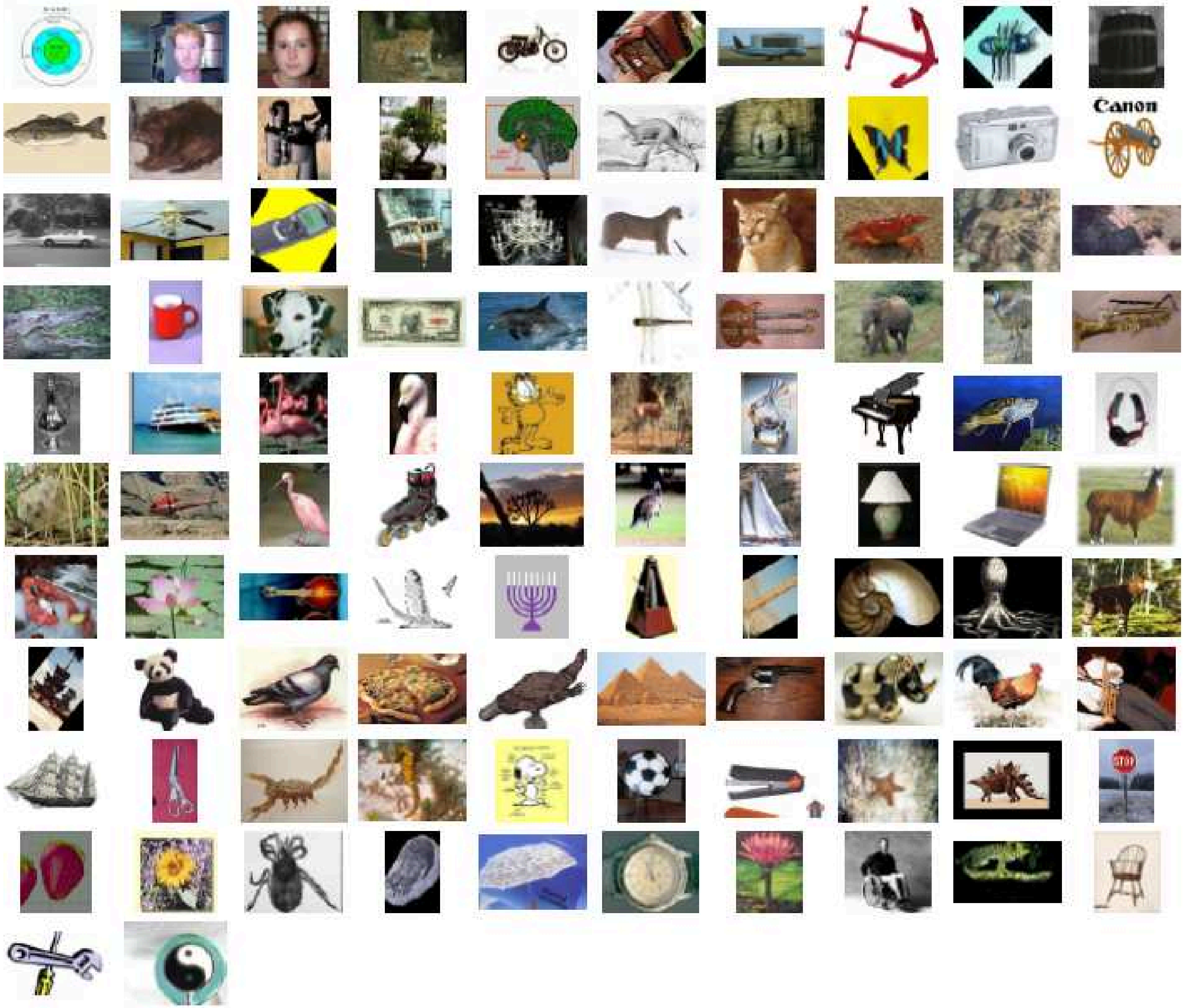
- Group Lasso
  - Block coordinate descent (Yuan and Lin, 2006)
  - Active set method (Roth and Fischer, 2008; Obozinski et al., 2009)
  - Proximal methods (Liu et al., 2009)

- MKL
  - Dual ascent, e.g., sequential minimal optimization (Bach et al., 2004a)
  - η-trick + cutting-planes (Sonnenburg et al., 2006)
  - η-trick + projected gradient descent (Rakotomamonjy et al., 2008)
  - Active set (Bach, 2008c)

Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**
- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)
- Two strategies for kernel combinations:
  - Uniform combination $\Leftrightarrow$ ℓ_2-norm
  - Sparse combination $\Leftrightarrow$ ℓ_1-norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - In particular, with few well-designed kernels
    - Be careful with normalization of kernels (Bach et al., 2004b)

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. all)
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		**77.72** $\pm$ **0.94**	**83.78** $\pm$ **0.39**
(Varma and Ray, 2007)		**81.54** $\pm$ **1.08**	**89.56** $\pm$ **0.59**

Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**
- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)
- Two strategies for kernel combinations:
  - Uniform combination $\Leftrightarrow$ ℓ_2-norm
  - Sparse combination $\Leftrightarrow$ ℓ_1-norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - In particular, with few well-designed kernels
    - Be careful with normalization of kernels (Bach et al., 2004b)

Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**

- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)

- Two strategies for kernel combinations:
  - Uniform combination $\Leftrightarrow$ ℓ_2-norm
  - Sparse combination $\Leftrightarrow$ ℓ_1-norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - In particular, with few well-designed kernels
    - Be careful with normalization of kernels (Bach et al., 2004b)

- **Sparse methods**: new possibilities and new features

Non-linear variable selection

- Given $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$, find function $f(x_1, \ldots, x_q)$ which depends only on a few variables

- Sparse generalized additive models (e.g., MKL):

  – restricted to $f(x_1, \ldots, x_q) = f_1(x_1) + \cdots + f_q(x_q)$

- Cosso (Lin and Zhang, 2006):

  – restricted to $f(x_1, \ldots, x_q) = \sum_{J \subset \{1,\ldots,q\},\ |J| \leqslant 2} f_J(x_J)$

Non-linear variable selection

- Given $x = (x_1, \dots, x_q) \in \mathbb{R}^q$, find function $f(x_1, \dots, x_q)$ which depends only on a few variables

- Sparse generalized additive models (e.g., MKL):
  - restricted to $f(x_1, \dots, x_q) = f_1(x_1) + \dots + f_q(x_q)$

- Cosso (Lin and Zhang, 2006):
  - restricted to $f(x_1, \dots, x_q) = \sum_{J \subset \{1,\dots,q\},\ |J| \leqslant 2} f_J(x_J)$

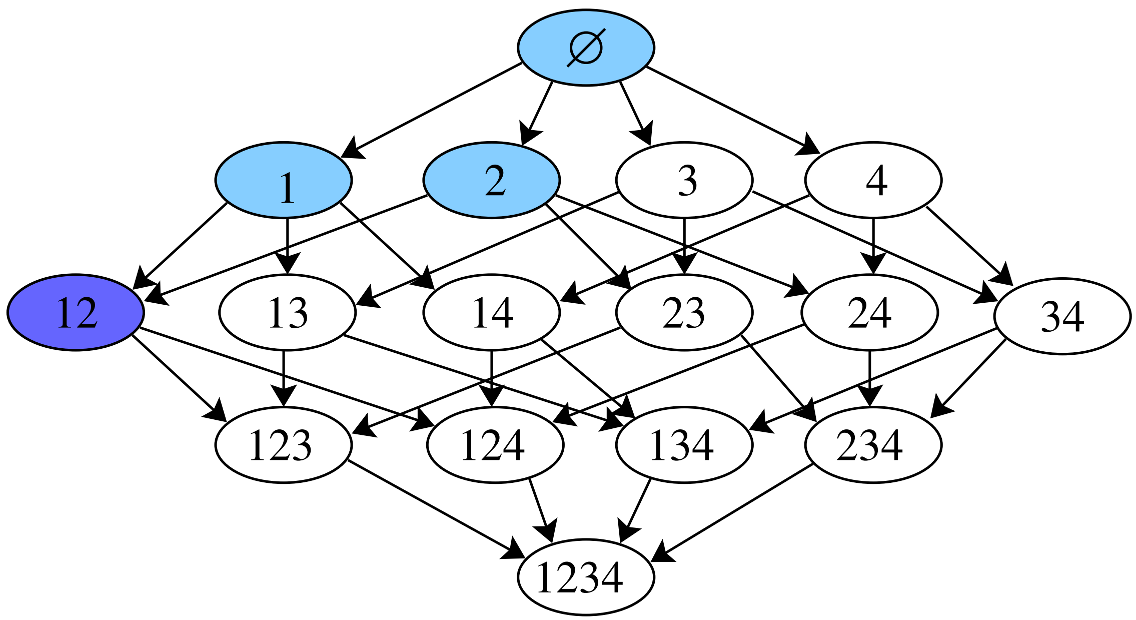
- Universally consistent non-linear selection requires all 2^q subsets

$$f(x_1, \dots, x_q) = \sum_{J \subset \{1,\dots,q\}} f_J(x_J)$$

Restricting the set of active kernels (Bach, 2008c)

- 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 $\{1, \dots, q\}$ with inclusion DAG
  - Select a subset only after all its subsets have been selected



DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

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

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

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

- **Hierarchical kernel learning** (Bach, 2008c) :
  - **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

Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Groups of features**
  - Non-linearity: Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)
- **Structured sparsity**
  - Overlapping groups and hierarchies

References

C. Archambeau and F. Bach. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2008.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008a.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b.

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

F. Bach. Self-concordant analysis for logistic regression. Technical Report 0910.4627, ArXiv, 2009.

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.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516, 2008.

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

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizbal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer-Verlag, 2000.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

E.J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37 (5A):2145–2177, 2009.

F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In *25th International Conference on Machine Learning (ICML)*, 2008.

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

D.L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452, 2005.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32 (2):407–451, 2004.

J. Fan and R. Li. Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1361, 2001.

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

J. Friedman, T. Hastie, H. H
"ofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2): 302–332, 2007.

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

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

J. Huang and T. Zhang. The benefit of group sparsity. Technical Report 0901.2962v2, ArXiv, 2009.

J. Huang, S. Ma, and C.H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009.

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. *Bioinformatics*, 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.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

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

J. Liu, S. Ji, and J. Ye. Multi-Task Feature Learning Via Efficient l2,-Norm Minimization. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528, 2009.

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

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. Technical report, arXiv: 0809.2932, 2008.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

R.M. Neal. *Bayesian learning for neural networks*. Springer Verlag, 1996.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Pub, 2003.

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

G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

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

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

V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning*

(ICML), 2008.

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

M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 9:759–813, 2008.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

S. Sonnenburg, G. Raetsch, C. Schaefer, and B. Schoelkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

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

S. A. Van De Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36 (2):614, 2008.

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. *IEEE transactions on information theory*, 55(5):2183, 2009.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

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.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 22, 2008a.

T. Zhang. Multi-stage convex relaxation for learning with sparse regularization. *Advances in Neural Information Processing Systems*, 22, 2008b.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *The Journal of Machine Learning Research*, 10:555–568, 2009.

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

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

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320, 2005.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.