Structured sparsity through convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with R. Jenatton, J. Mairal, G. Obozinski

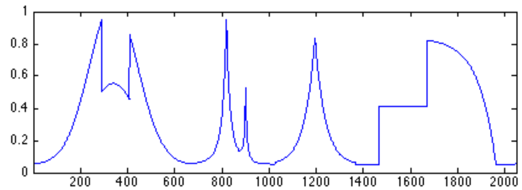
IRISA - October 2012

Outline

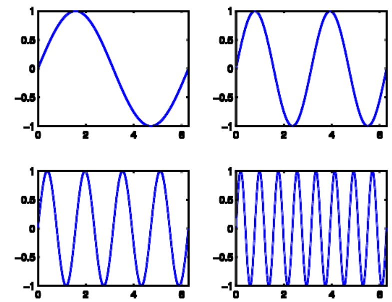
- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices
- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

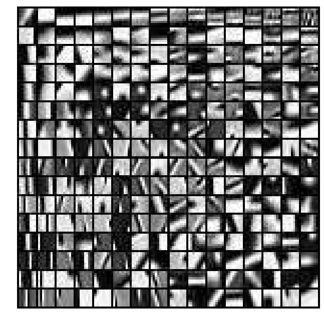
Sparsity in signal processing

- Let $x \in \mathbb{R}^m$ be a signal



- Let $D = [d_1, \ldots, d_p] \in \mathbb{R}^{m \times p}$ be a set of "basis vectors". D $=$ **dictionary**





- D is "adapted" to x if it can represent it with a few basis vectors:
  - there exists a sparse vector α in $\mathbb{R}^p$ such that $x \approx D\alpha$.
    $\alpha =$ **sparse code**

$$\underbrace{\begin{pmatrix} \\ x \\ \\ \end{pmatrix}}_{x \in \mathbb{R}^m} \approx \underbrace{\begin{pmatrix} & & & \\ d_1 & d_2 & \cdots & d_p \\ & & & \end{pmatrix}}_{D \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}}_{\alpha \in \mathbb{R}^p,\ \text{sparse}}$$

Sparsity in signal processing
Sparse decomposition problem

$$\min_{\alpha \in \mathbb{R}^p} \underbrace{\frac{1}{2}||x - D\alpha||_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\alpha)}_{\substack{\text{sparsity-inducing} \\ \text{regularization}}}$$

- The term ψ induces sparsity
  - the ℓ_0 "pseudo-norm": $||\alpha||_0 \triangleq \#\{i \text{ s.t. } \alpha_i \neq 0\}$ (NP-hard)
  - the ℓ_1 norm: $||\alpha||_1 \triangleq \sum_{i=1}^{p} |\alpha_i|$ (convex)
  - ...

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a linear function $w^\top \Phi(x)$ of features $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$
- **(regularized) empirical risk minimization**: find $\hat{w}$ solution of

$$\min_{w \in \mathcal{F}} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, w^\top \Phi(x_i)\big) \quad + \quad \mu \Omega(w)$$

convex data fitting term + regularizer

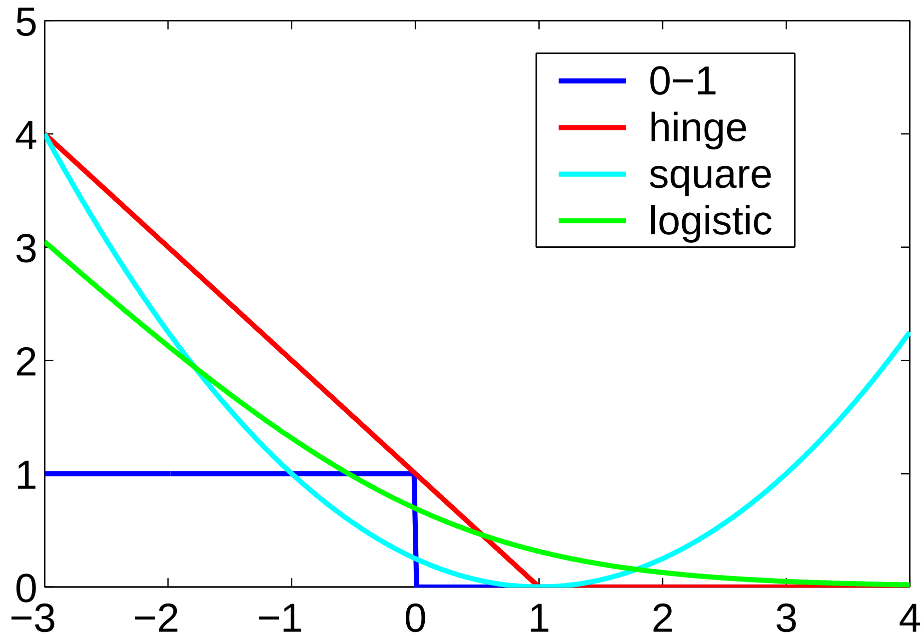
Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = w^\top \Phi(x)$
  - quadratic loss $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - w^\top \Phi(x))^2$

Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = w^\top \Phi(x)$
  - quadratic loss $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - w^\top \Phi(x))^2$

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



Usual regularizers

- **Goal**: avoid overfitting

- **(squared) Euclidean norm**: $\|w\|_2^2 = \sum_{j=1}^{p} |w_j|^2$
  - Numerically well-behaved
  - Representer theorem and kernel methods : $w = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)

Usual regularizers

- **Goal**: avoid overfitting

- **(squared) Euclidean norm**: $\|w\|_2^2 = \sum_{j=1}^p |w_j|^2$
  - Numerically well-behaved
  - Representer theorem and kernel methods : $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)

- **Sparsity-inducing norms**
  - Main example: ℓ_1-norm $\|w\|_1 = \sum_{j=1}^p |w_j|$
  - Perform model selection as well as regularization
  - Non-smooth optimization and structured sparsity
  - See, e.g., Bach, Jenatton, Mairal, and Obozinski (2011)

Sparsity in supervised machine learning

- Observed data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
  - Response vector $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$
  - Design matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$
- Regularized empirical risk minimization:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w) = \boxed{\min_{w \in \mathbb{R}^p} \; L(y, Xw) + \lambda \Omega(w)}$$

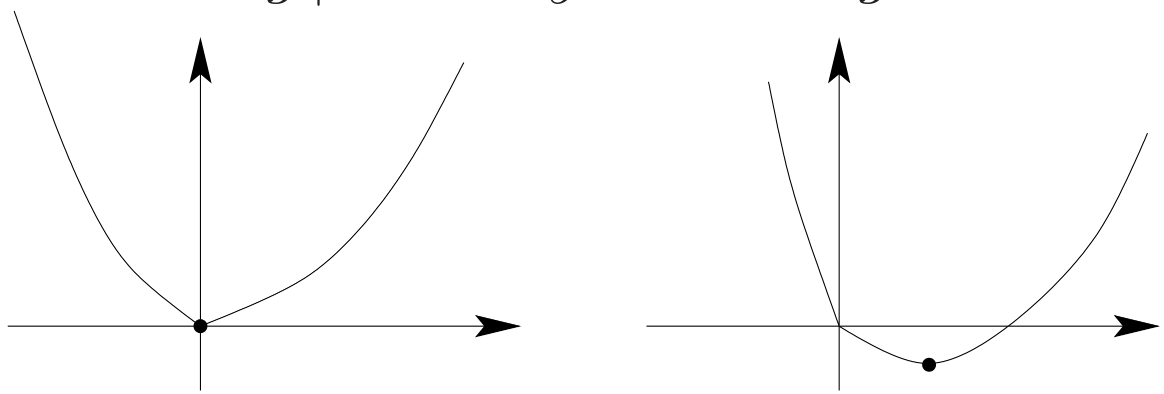
- Norm Ω to promote sparsity
  - square loss + ℓ_1-norm $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$

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?

Why ℓ_1-norms lead to sparsity?

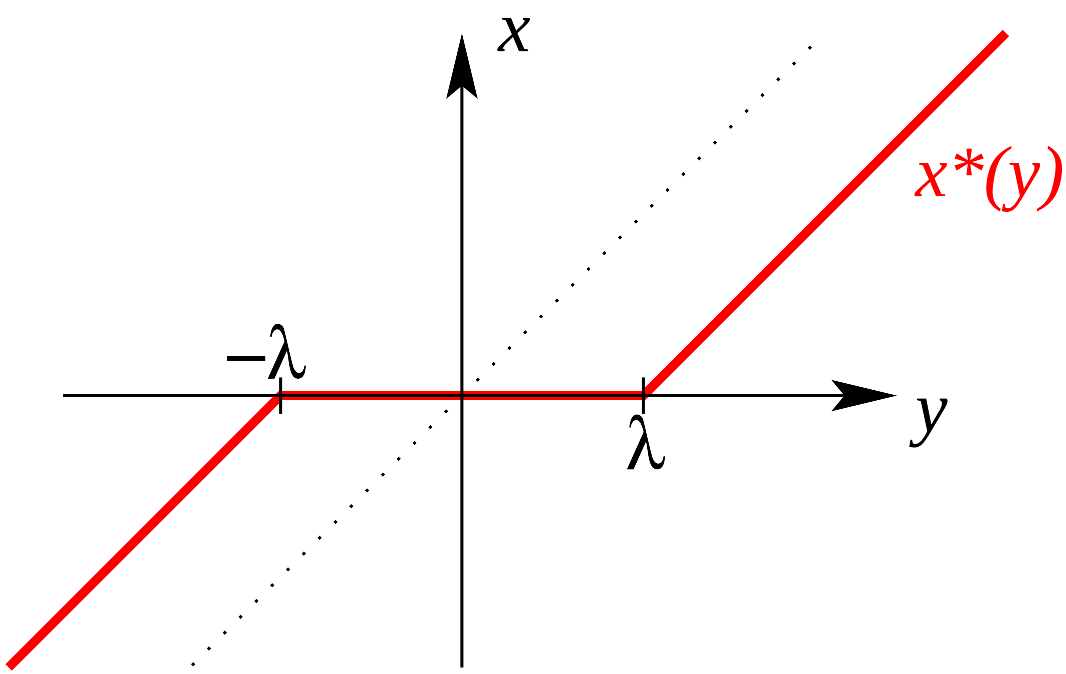
- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x\in\mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$



  - $x = 0$ 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

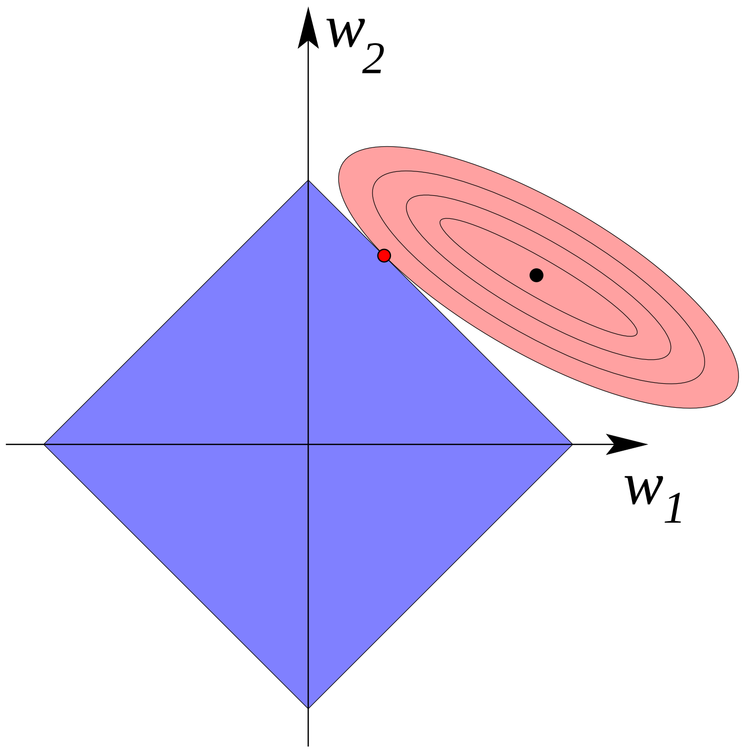
Why ℓ_1-norms lead to sparsity?

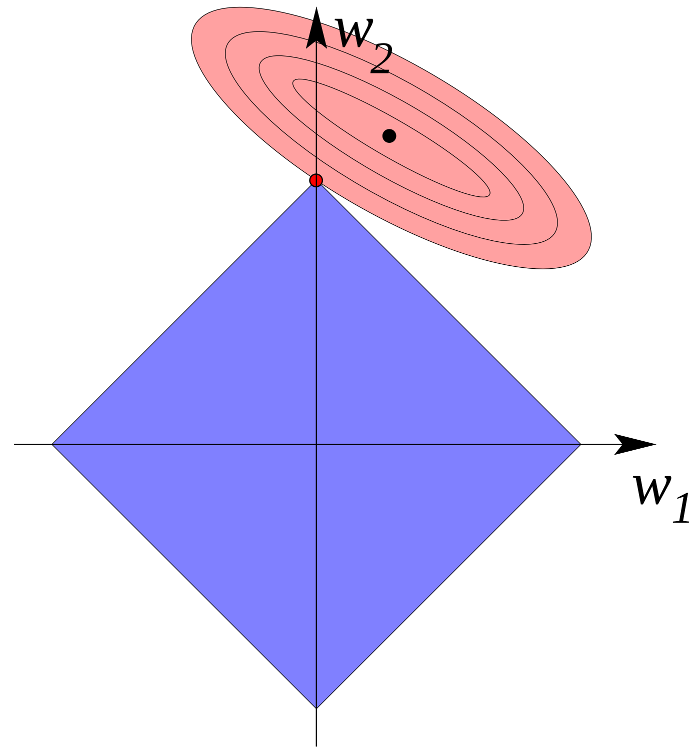
- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
- Solution $\boxed{x^* = \operatorname{sign}(y)(|y| - \lambda)_+}$ = soft thresholding



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





A review of nonsmooth convex analysis and optimization

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

A review of nonsmooth convex analysis and optimization

- Analysis: optimality conditions
  - Convex duality
- Optimization: algorithms
  - First-order methods
- **Books**: Boyd and Vandenberghe (2004), Bonnans et al. (2003), Bertsekas (1995), Borwein and Lewis (2000), Nesterov (2003)
- **Simple techniques might not work!**

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)

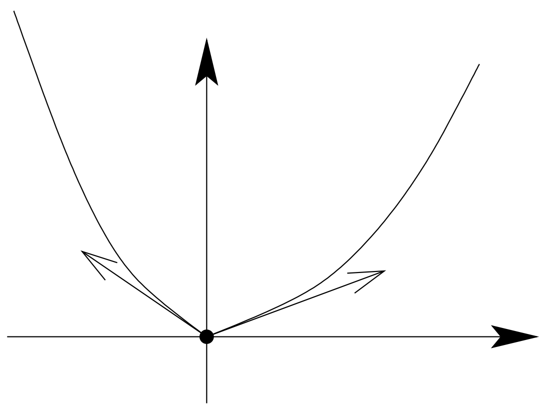
Directional derivatives - convex functions on $\mathbb{R}^p$

- 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 $j \in \{1, \ldots, p\}$,

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

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

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

  – For $J \subset \{1, \ldots, p\}$, $X_J \in \mathbb{R}^{n \times |J|} = X(:, J)$ 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

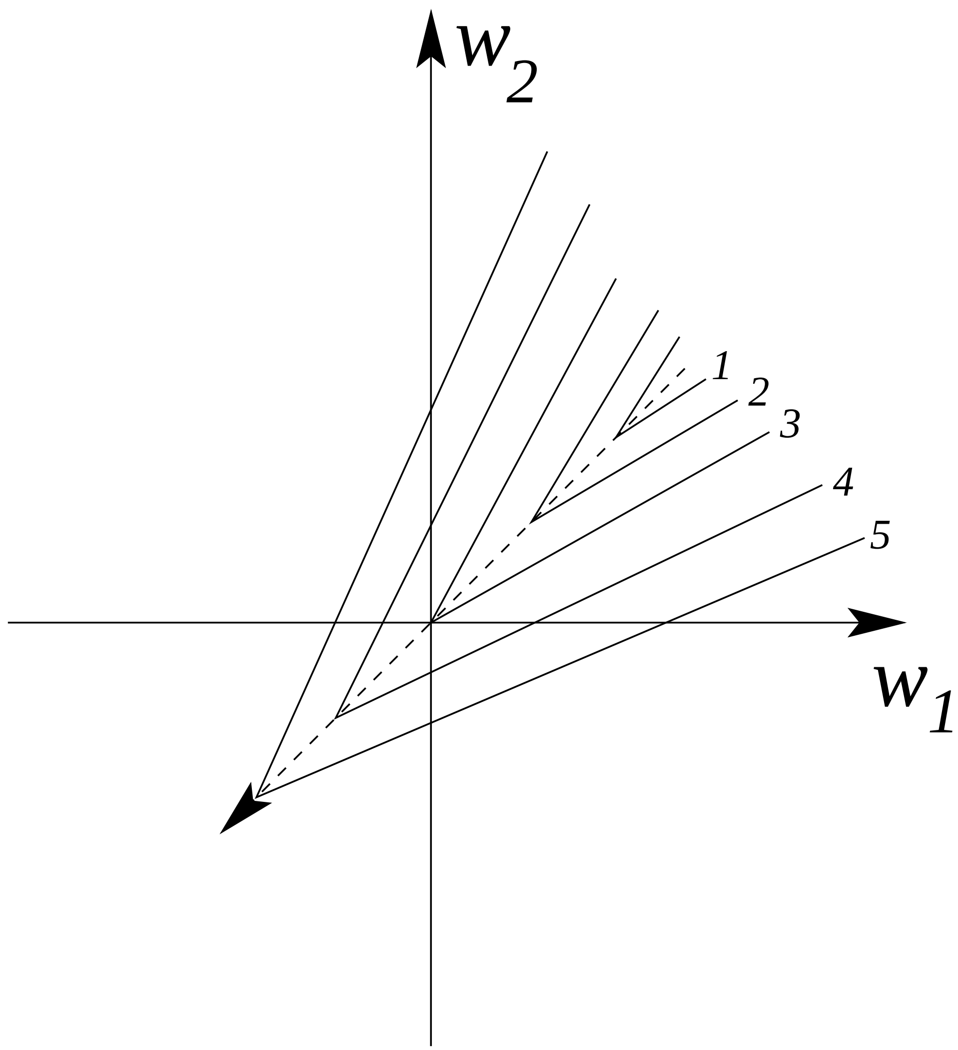
- **Gradient descent**: $w_{t+1} = w_t - \alpha_t \nabla J(w_t)$
  - with line search: search for a decent (not necessarily best) α_t
  - fixed diminishing step size, e.g., $\alpha_t = a(t+b)^{-1}$

- 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)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{\mu}{2} \|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{\mu} \nabla L(w_t)$

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 $w_{t+1} = \mathrm{SoftThres}(w_t - \frac{1}{\mu}\nabla L(w_t))$

- 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., 2009b)
  - 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., 2009b)
  - 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 $\Rightarrow$ 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: $\mathrm{sign}(w_J) = s_J$
  - 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,

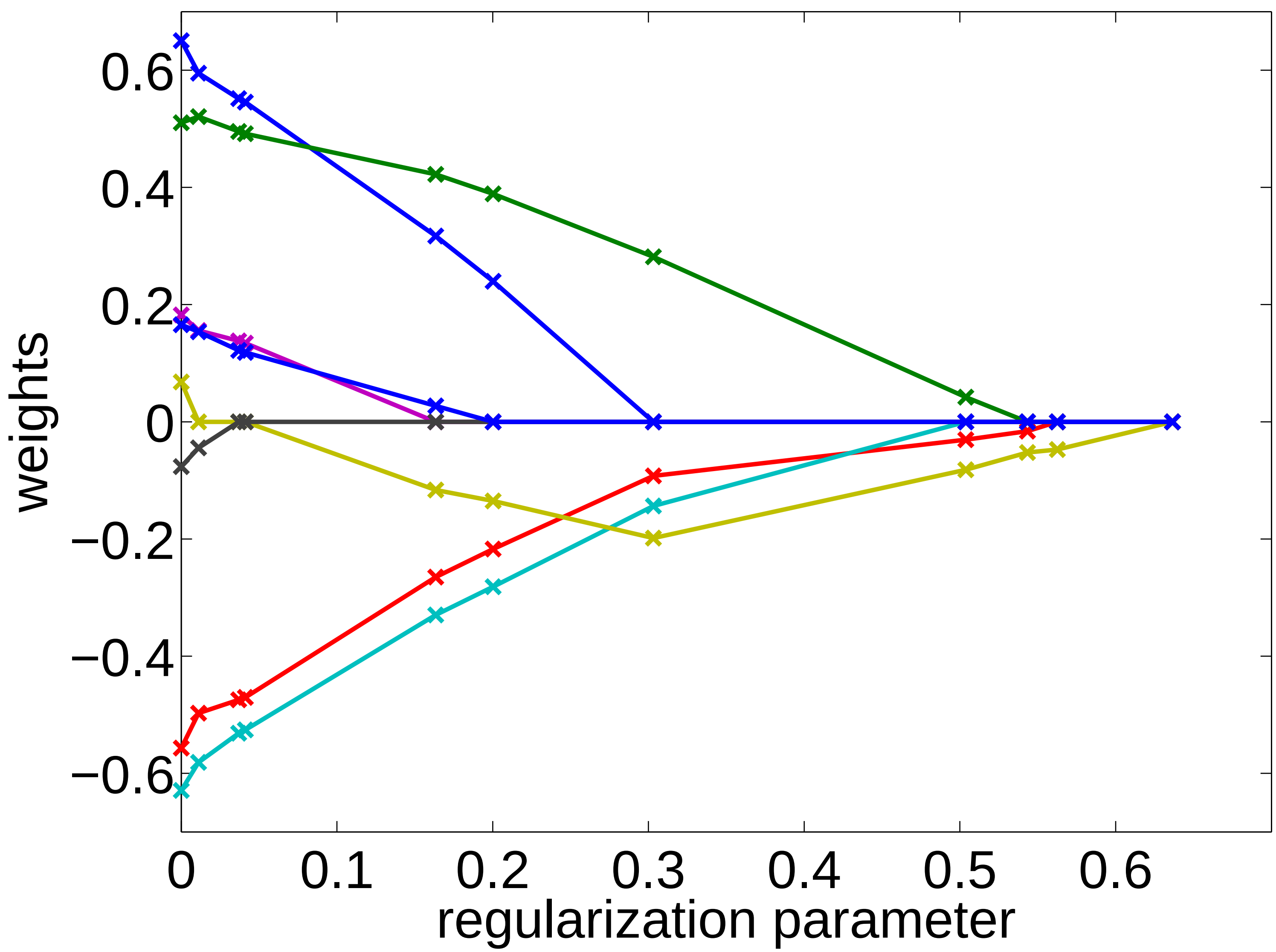
$$w_J^*(\lambda) = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$$

  valid, as long as,

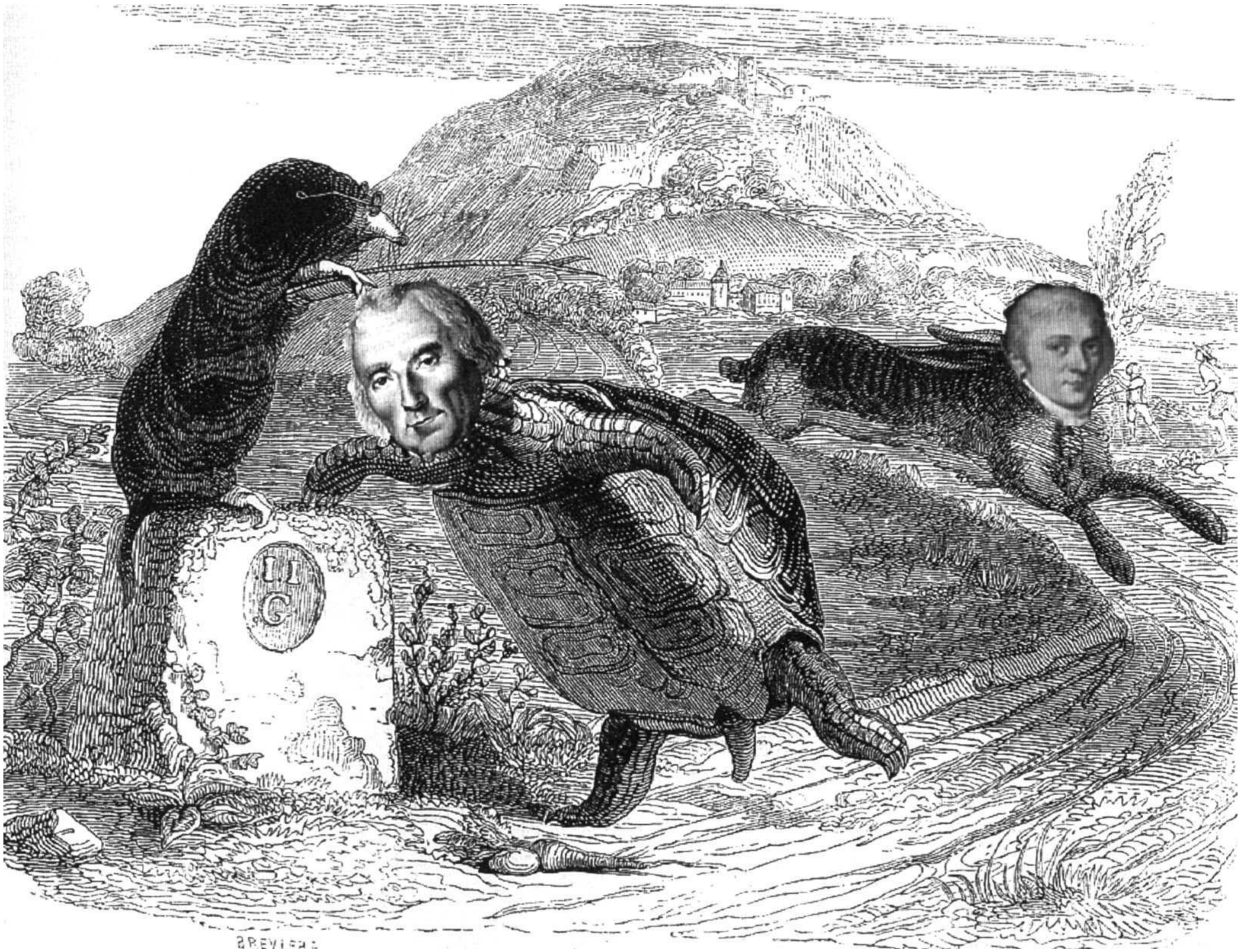
  – sign condition: $\mathrm{sign}(w_J^*(\lambda)) = s_J$
  – subgradient condition: $\|X_{J^c}^\top (X_J w_J^*(\lambda) - y)\|_\infty \leqslant \lambda$
  – this defines an interval on λ: the path is thus **piecewise affine**

- Simply need to find break points and directions

Piecewise linear paths



Gaussian hare vs. Laplacian tortoise

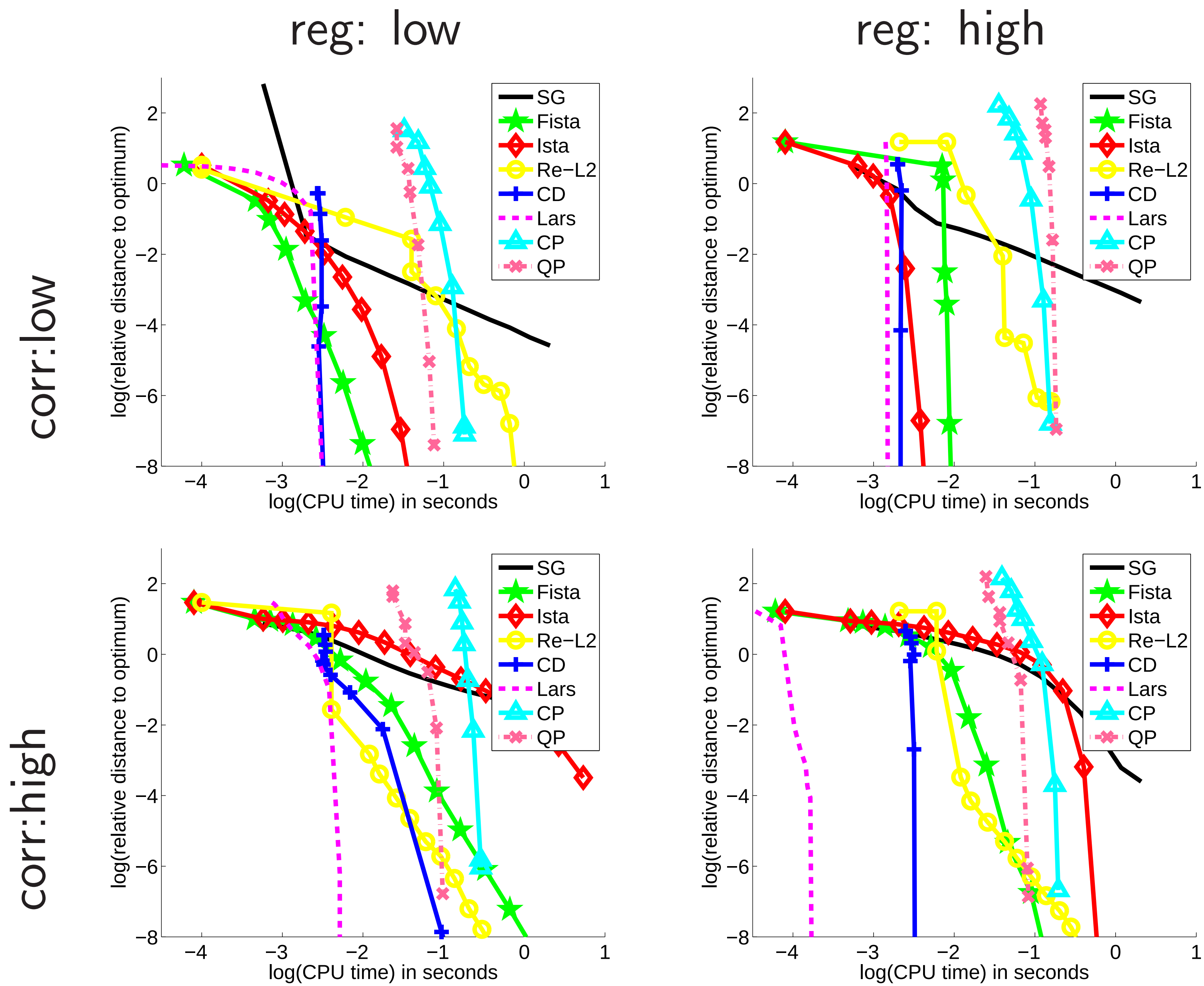


- Coord. descent and proximal: $O(pn)$ per iterations for ℓ_1 and ℓ_2
- “Exact” algorithms: $O(kpn)$ for ℓ_1 **vs.** $O(p^2n)$ for ℓ_2

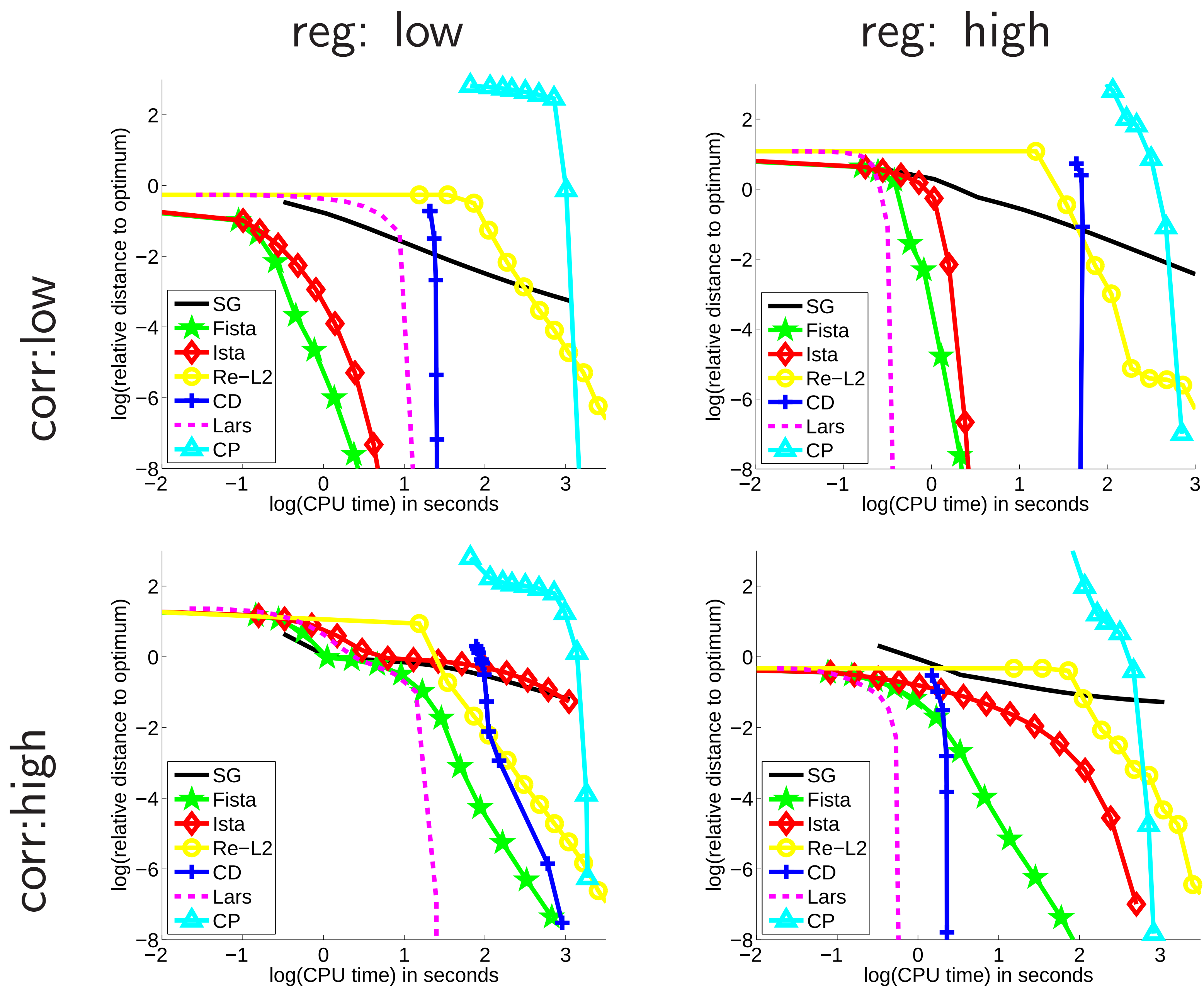
Additional methods - Softwares

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

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

Outline

- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices

- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications

- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

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

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

  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 $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

- Condition depends on $\mathbf{w}$ and $\mathbf{J}$ (may be relaxed)
  - may be relaxed by maximizing out $\mathrm{sign}(\mathbf{w})$ 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 $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

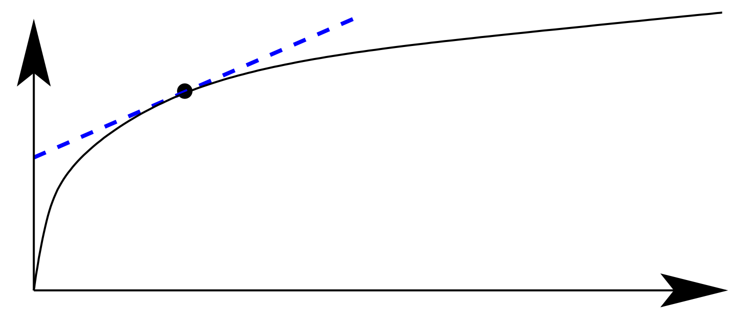
- **The Lasso is usually not model-consistent**
  - Selects more variables than necessary (see, e.g., Lv and Fan, 2009)
  - **Fixing the Lasso**: adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2008), thresholding (Lounici, 2008), Bolasso (Bach, 2008a), stability selection (Meinshausen and Bühlmann,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

Adaptive Lasso and concave penalization

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

- **Reformulation in terms of concave penalization**

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

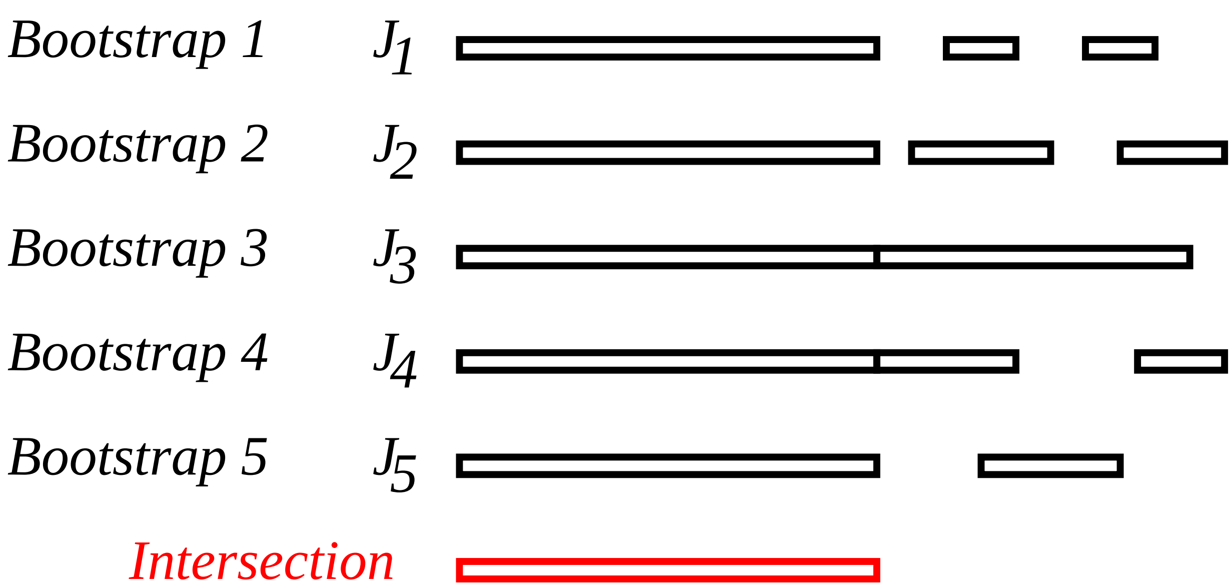


  - Example: $g(|w_j|) = |w_j|^{1/2}$ or $\log |w_j|$. Closer to the ℓ_0 penalty
  - Concave-convex procedure: replace $g(|w_j|)$ by affine upper bound
  - Better sparsity-inducing properties (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2008b)

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

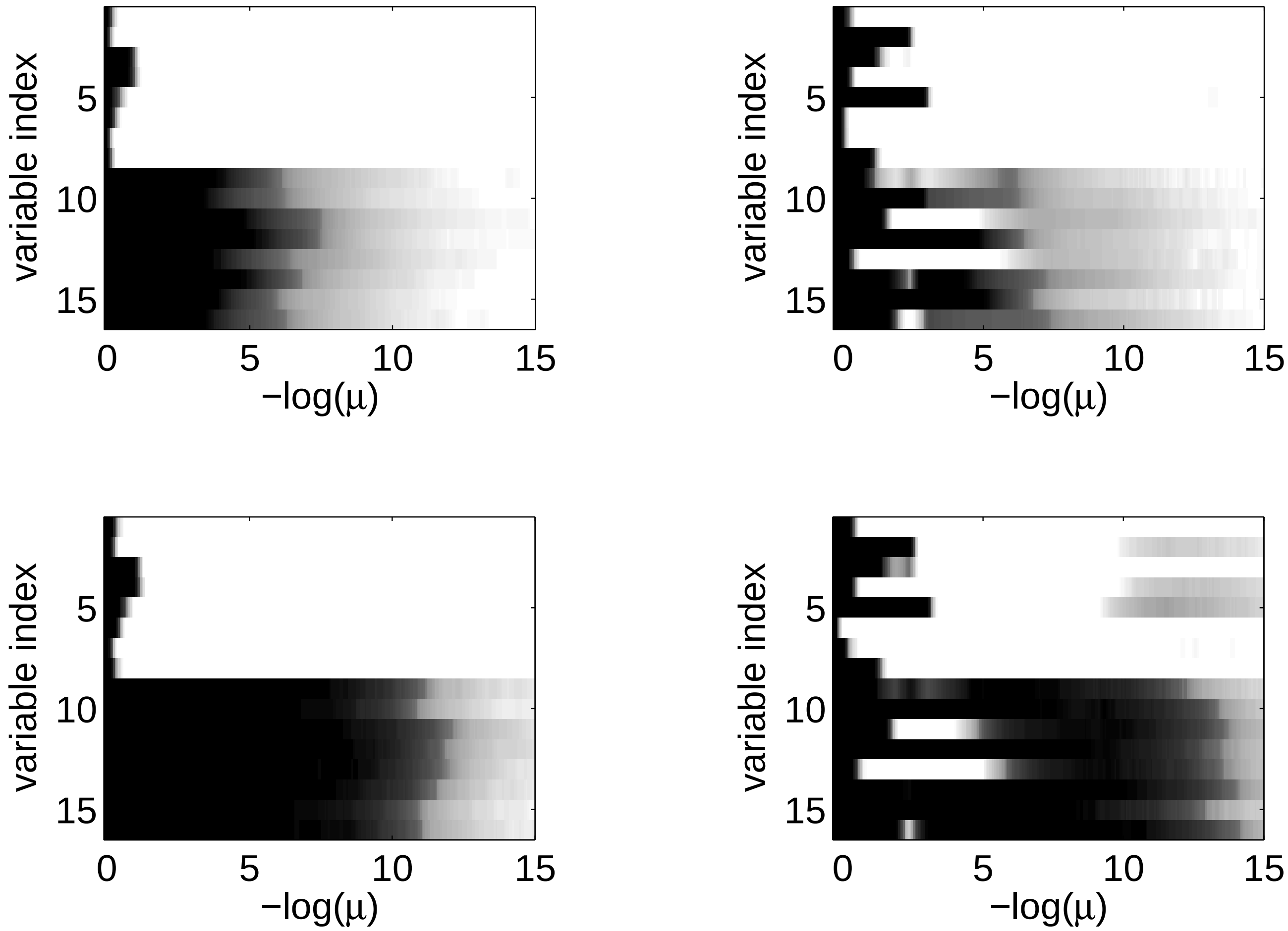


Model selection consistency of the Lasso/Bolasso

- 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)

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

- **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
  - **<u>and</u>** 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)
- **Slow rate for predictions if no assumptions**: $\sqrt{\frac{k \log p}{n}}$

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 $\min\limits_{j, \mathbf{w}_j \neq 0} |\mathbf{w}_j| > C\sigma\sqrt{\frac{\log p}{n}}$

- 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

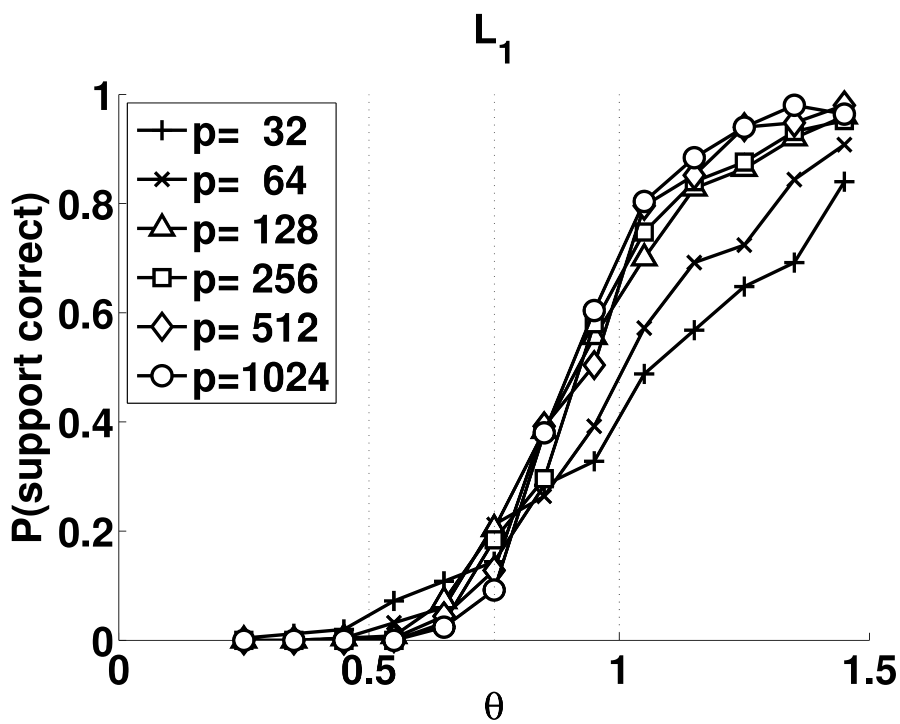
$$\text{estimation error} \qquad \|\hat{w} - \mathbf{w}\|_1 \leqslant \frac{16A}{\kappa^2(k)} \sigma k \sqrt{\frac{\log p}{n}}$$

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

- 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

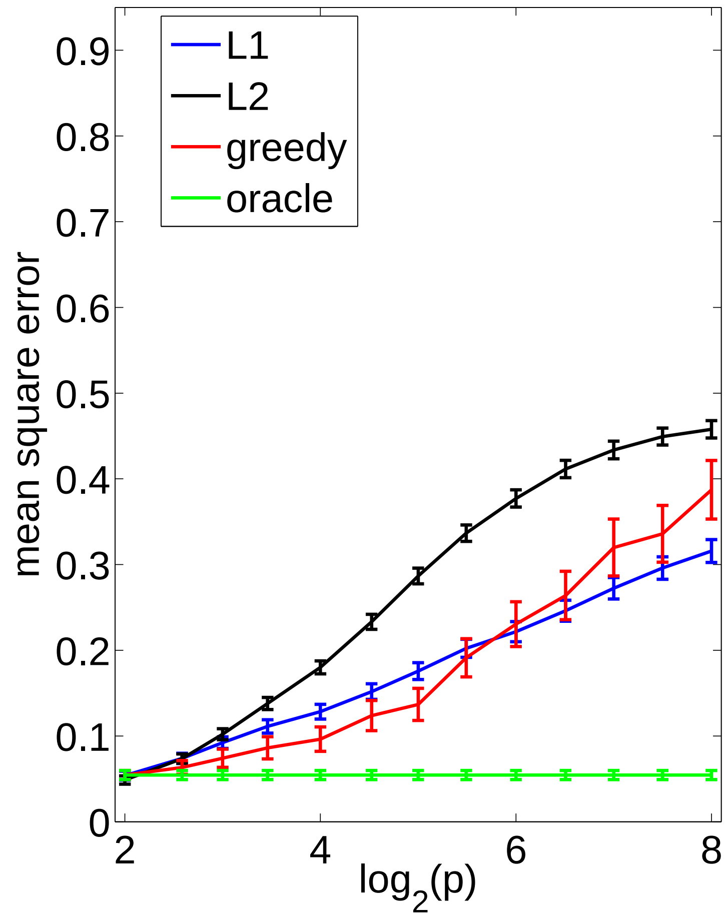
- Compared methods to reach the least-square solution
  - Ridge regression: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
  - Lasso: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$
  - Forward greedy:
    - Initialization 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

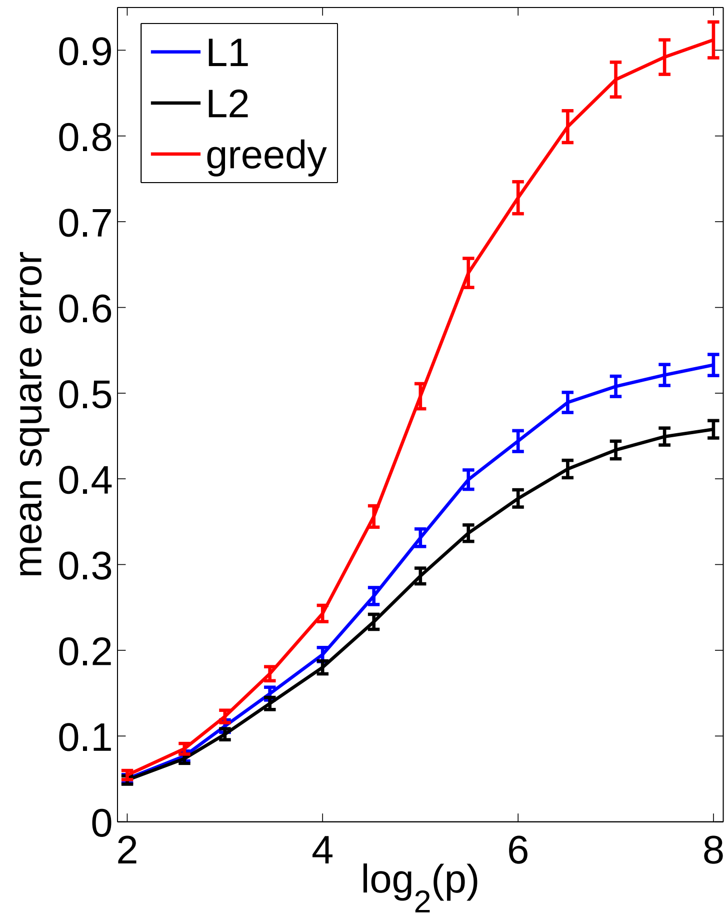
- Regularization parameters selected on the test set

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 (next part of the tutorial)
- **Sparse methods are not limited to variable selection in a linear model**
  - Multiple kernel learning

Going beyond the Lasso
Non-linearity - Multiple kernel learning

- **Multiple kernel learning**
  - Learn sparse combination of matrices $k(x,x') = \sum_{j=1}^p \eta_j k_j(x,x')$
  - Mixing positive aspects of ℓ_1-norms and ℓ_2-norms

- **Equivalent to group Lasso**
  - p multi-dimensional features $\Phi_j(x)$, where

$$k_j(x,x') = \Phi_j(x)^\top \Phi_j(x')$$

  - learn predictor $\sum_{j=1}^p w_j^\top \Phi_j(x)$
  - Penalization by $\sum_{j=1}^p \|w_j\|_2$ (Bach et al., 2004)

Going beyond the Lasso
Structured set of features

- **Dealing with exponentially many features**
  - Can we design efficient algorithms for the case $\log p \approx n$?
  - Use structure to reduce the number of allowed patterns of zeros
  - Recursivity, **hierarchies** and factorization
- **Prior information on sparsity patterns**
  - Grouped variables with overlapping groups

Outline

- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices
- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

Learning on matrices - Image denoising

- Simultaneously denoise all patches of a given image
- Example from Mairal, Bach, Ponce, Sapiro, and Zisserman (2009e)

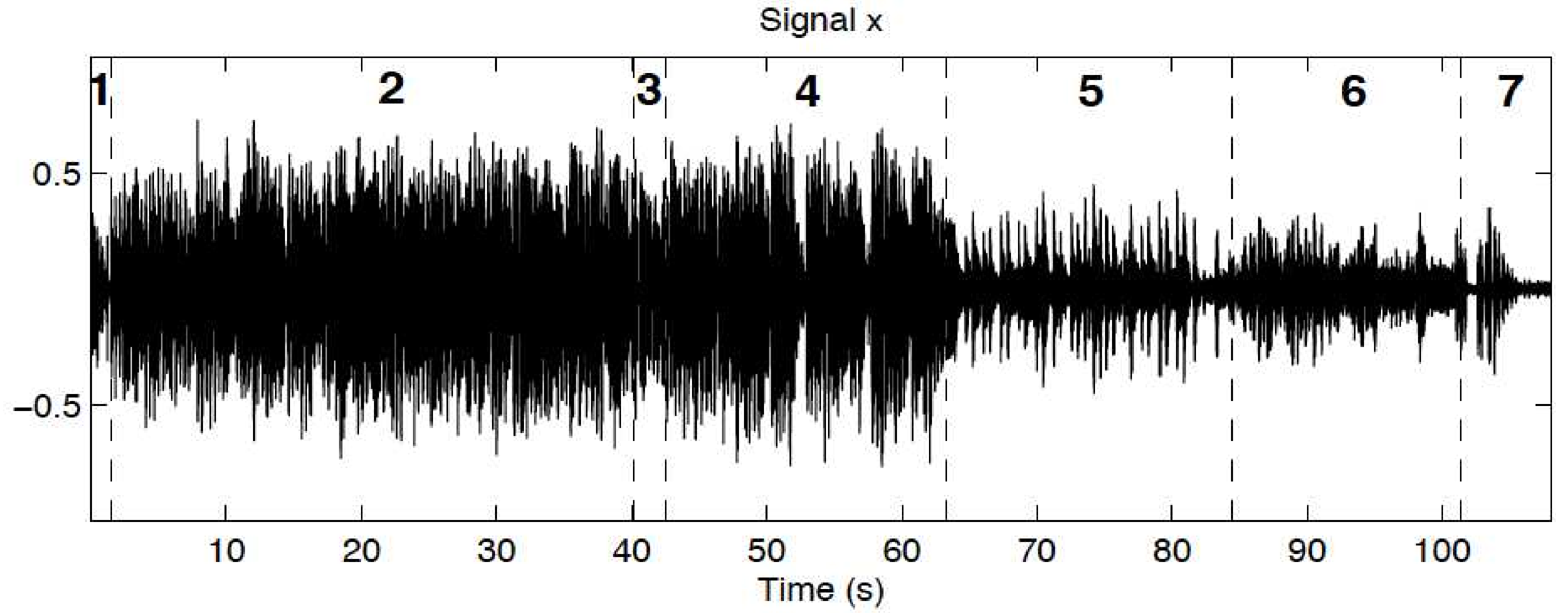
Learning on matrices - Collaborative filtering

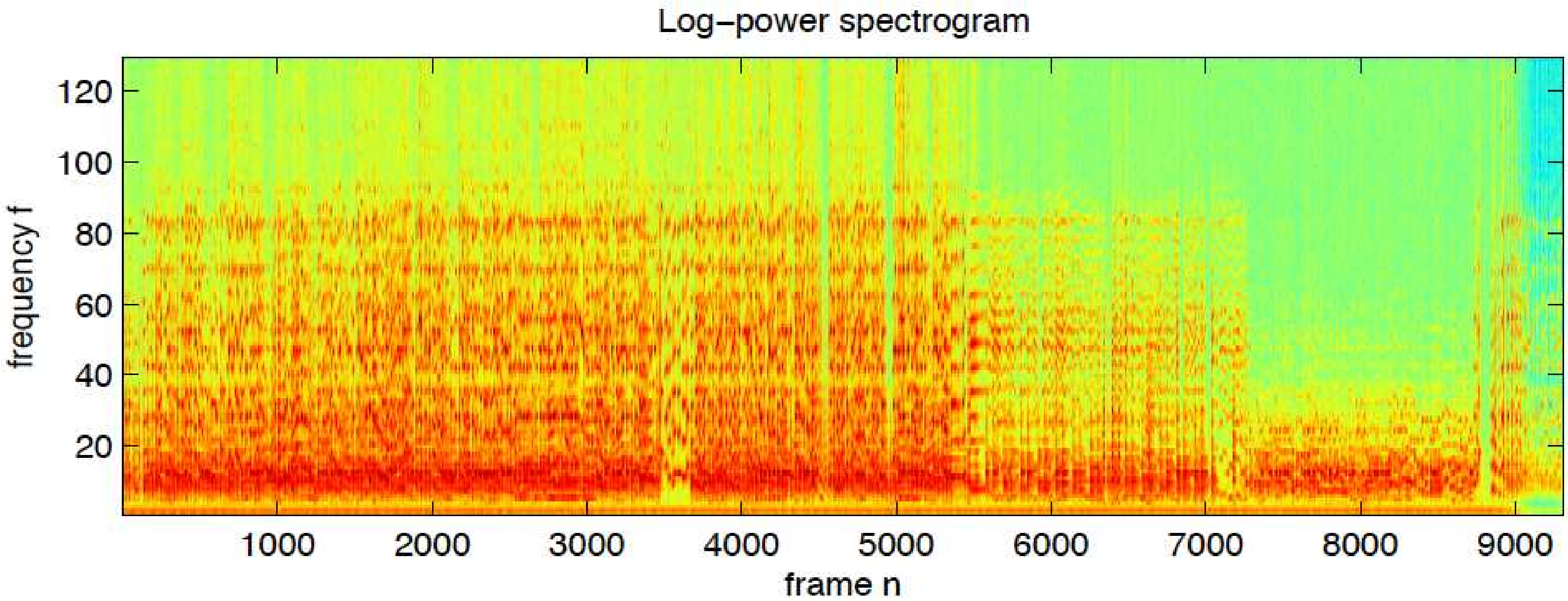
- Given $n_{\mathcal{X}}$ "movies" $\mathbf{x} \in \mathcal{X}$ and $n_{\mathcal{Y}}$ "customers" $\mathbf{y} \in \mathcal{Y}$,
- predict the "rating" $z(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ of customer $\mathbf{y}$ for movie $\mathbf{x}$
- Training data: large $n_{\mathcal{X}} \times n_{\mathcal{Y}}$ incomplete matrix $\mathbf{Z}$ that describes the known ratings of some customers for some movies
- **Goal**: complete the matrix.

	1				1								2							
											2				1				3	
	2									3						3				3
						1			2					1				1		
		1		3							1						1	3		
	1		2													2				
								3												

Learning on matrices - Source separation

- Single microphone (Benaroya et al., 2006; Févotte et al., 2009)



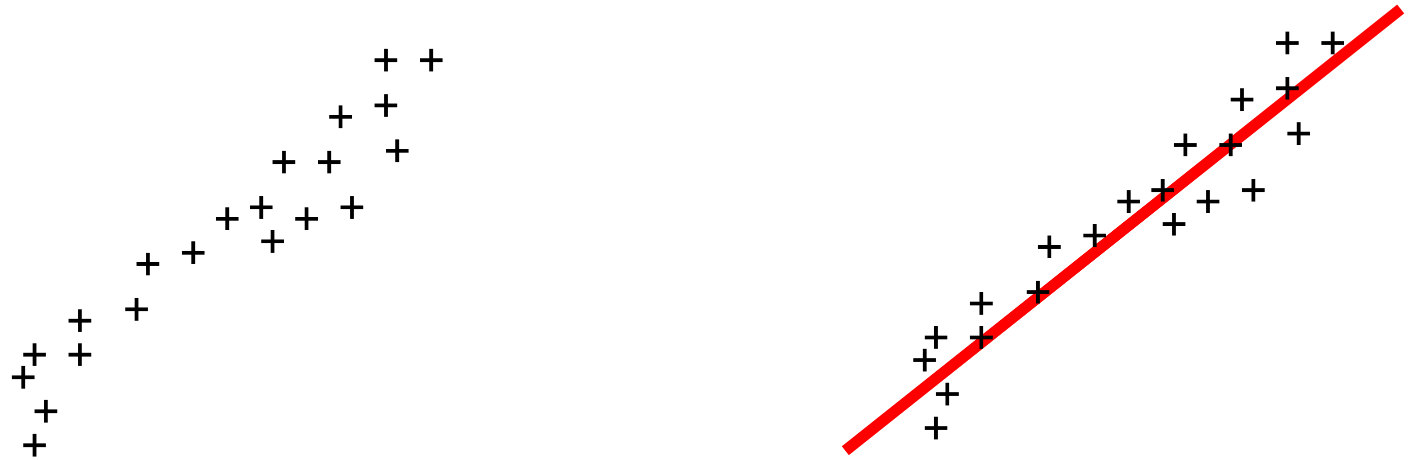


Learning on matrices - Multi-task learning

- k linear prediction tasks on same covariates $\mathbf{x} \in \mathbb{R}^p$
  - k weight vectors $\mathbf{w}_j \in \mathbb{R}^p$
  - Joint matrix of predictors $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$
- Classical application
  - Multi-category classification (one task per class) (Amit et al., 2007)
- **Share parameters between tasks**
- **Joint variable selection** (Obozinski et al., 2009)
  - Select variables which are predictive for all tasks
- **Joint feature selection** (Pontil et al., 2007)
  - Construct linear features common to all tasks

Matrix factorization - Dimension reduction

- Given data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$
  - **Principal component analysis**: $\boxed{\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$

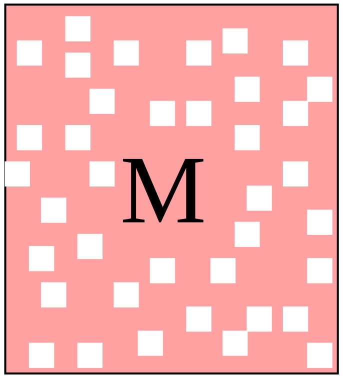


  - **K-means**: $\boxed{\mathbf{x}_i \approx \mathbf{d}_k \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$

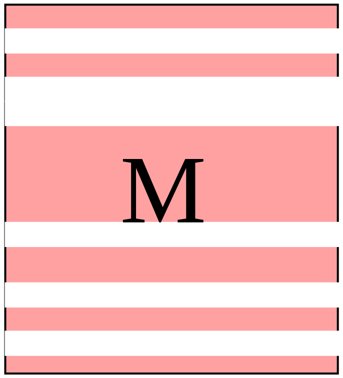
Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$

I - Directly on the elements of $\mathbf{M}$

- Many zero elements: $\mathbf{M}_{ij} = 0$



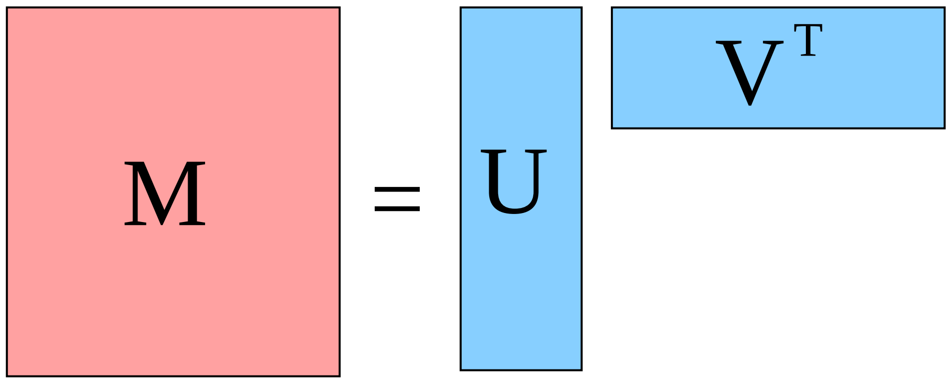
- Many zero rows (or columns): $(\mathbf{M}_{i1}, \ldots, \mathbf{M}_{ip}) = 0$



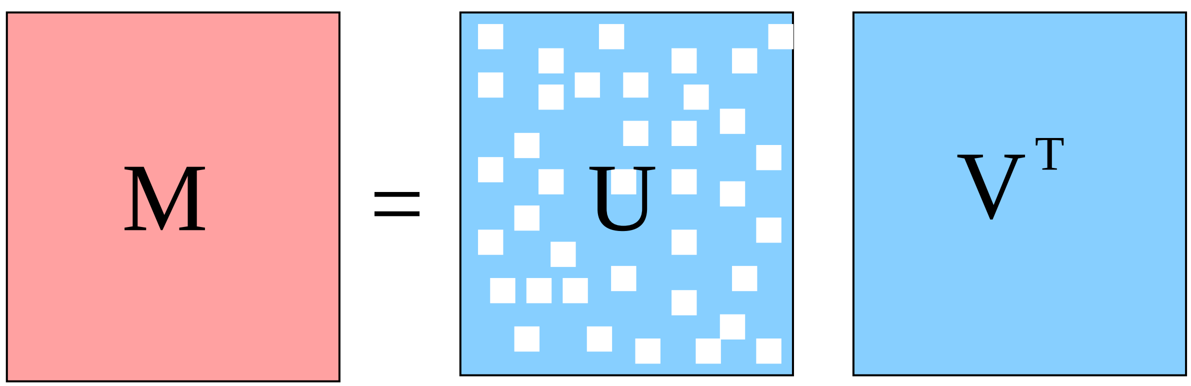
Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$

II - Through a factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$
- **Low rank**: m small



- **Sparse decomposition**: $\mathbf{U}$ sparse



Structured sparse matrix factorizations

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$

- **Structure on $\mathbf{U}$ and/or $\mathbf{V}$**
  - Low-rank: $\mathbf{U}$ and $\mathbf{V}$ have few columns
  - Dictionary learning / sparse PCA: $\mathbf{U}$ has many zeros
  - Clustering (k-means): $\mathbf{U} \in \{0,1\}^{n \times m}$, $\mathbf{U}\mathbf{1} = \mathbf{1}$
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros (Jenatton et al., 2010)
  - Low-rank + sparse (Candès et al., 2009)
  - etc.

- **Many applications**

- **Many open questions** (Algorithms, identifiability, etc.)

Multi-task learning

- Joint matrix of predictors $W = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k}$
- **Joint variable selection** (Obozinski et al., 2009)
  - Penalize by the sum of the norms of rows of W (group Lasso)
  - Select variables which are predictive for all tasks

Multi-task learning

- Joint matrix of predictors $W = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k}$

- **Joint variable selection** (Obozinski et al., 2009)
  - Penalize by the sum of the norms of rows of W (group Lasso)
  - Select variables which are predictive for all tasks

- **Joint feature selection** (Pontil et al., 2007)
  - Penalize by the trace-norm (see later)
  - Construct linear features common to all tasks

- Theory: allows number of observations which is sublinear in the number of tasks (Obozinski et al., 2008; Lounici et al., 2009)

- Practice: more interpretable models, slightly improved performance

Low-rank matrix factorizations

Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values

- Rank of $\mathbf{M}$ equal to the number of non-zero singular values

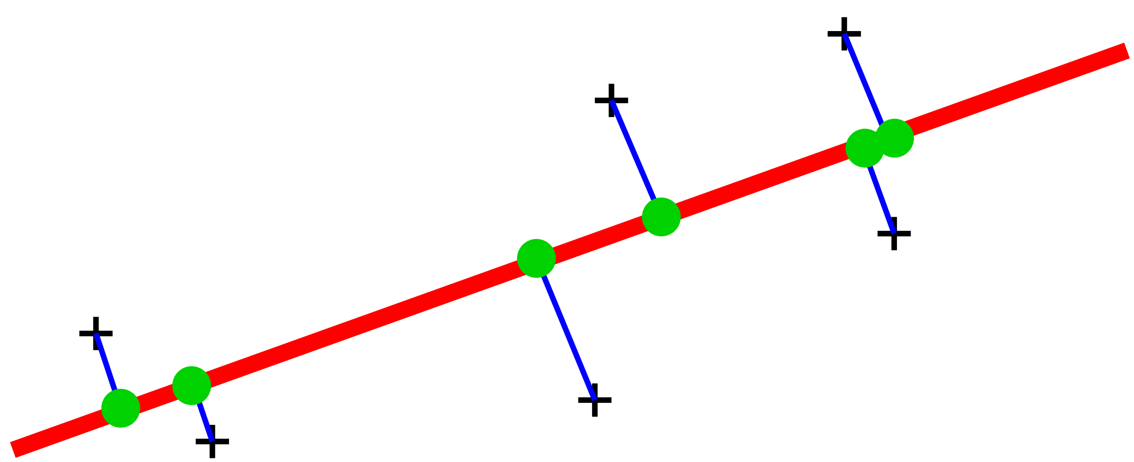
Low-rank matrix factorizations
Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values
- Rank of $\mathbf{M}$ equal to the number of non-zero singular values
- **Trace-norm (a.k.a. nuclear norm)** $=$ sum of singular values
- Convex function, leads to a semi-definite program (Fazel et al., 2001)
- First used for collaborative filtering (Srebro et al., 2005)
- Multi-category classif. (Amit et al., 2007; Harchaoui et al., 2012)

Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \ldots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent



Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \dots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent

- **Sparse extensions**
  - Interpretability
  - High-dimensional inference
  - Two views are differents
    - For analysis view, see d'Aspremont, Bach, and El Ghaoui (2008)

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

  – Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \Big\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \Big\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \big\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \big\|_2^2 \text{ is small}$$

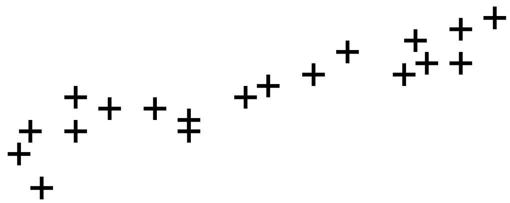
  - Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

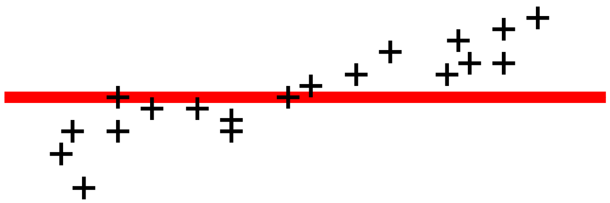
- Sparse formulation (Witten et al., 2009; Bach et al., 2008)
  - Penalize/constrain $\mathbf{d}_j$ by the ℓ_1-norm for sparsity
  - Penalize/constrain $\boldsymbol{\alpha}_i$ by the ℓ_2-norm to avoid trivial solutions

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_2 \leqslant 1$$

Sparse PCA vs. dictionary learning

- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse

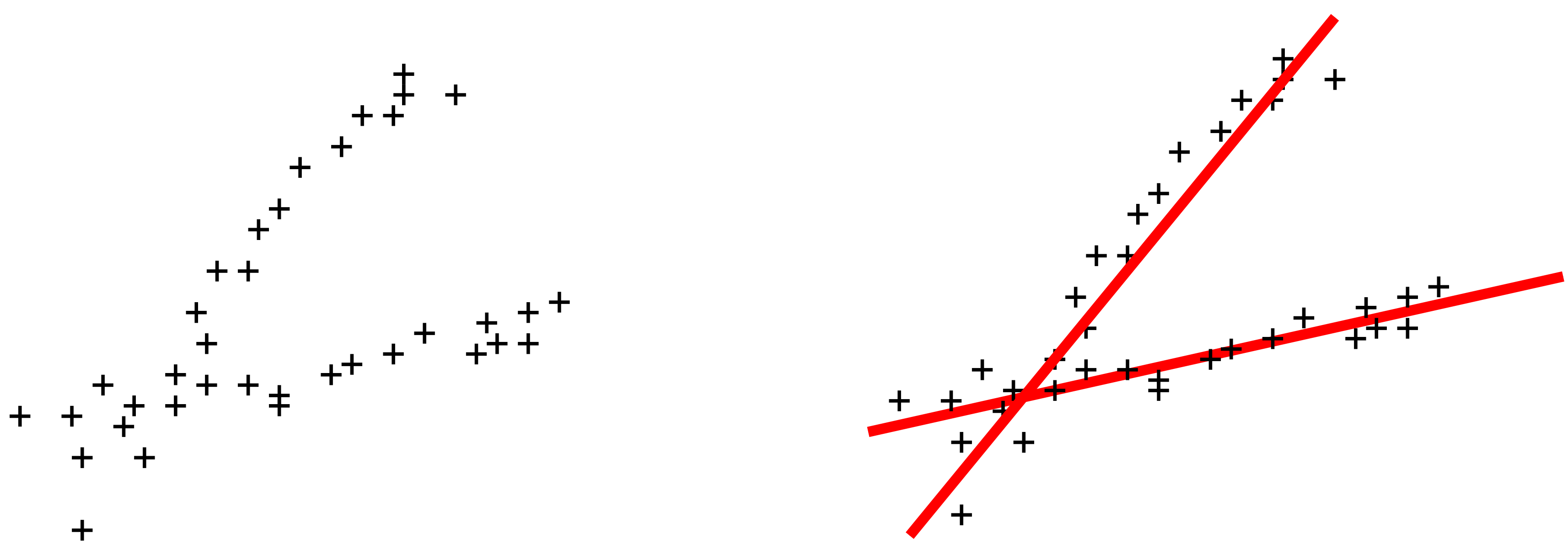




Sparse PCA vs. dictionary learning

- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse

- **Dictionary learning**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}_i$ sparse



Structured matrix factorizations (Bach et al., 2008)

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_\star \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_\bullet \leqslant 1$$

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_\bullet \text{ s.t. } \forall j, \|\mathbf{d}_j\|_\star \leqslant 1$$

- Optimization by alternating minimization (non-convex)
- $\boldsymbol{\alpha}_i$ decomposition coefficients (or "code"), $\mathbf{d}_j$ dictionary elements
- Two related/equivalent problems:
  - **Sparse PCA = sparse dictionary** (ℓ_1-norm on $\mathbf{d}_j$)
  - **Dictionary learning = sparse decompositions** (ℓ_1-norm on $\boldsymbol{\alpha}_i$) (Olshausen and Field, 1997; Elad and Aharon, 2006; Lee et al., 2007)

Dictionary learning for image denoising

$$\underbrace{\mathbf{x}}_{\text{measurements}} = \underbrace{\mathbf{y}}_{\text{original image}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{noise}}$$

Dictionary learning for image denoising

- **Solving the denoising problem** (Elad and Aharon, 2006)
  - Extract all overlapping 8×8 patches $\mathbf{x}_i \in \mathbb{R}^{64}$
  - Form the matrix $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{n \times 64}$
  - Solve a matrix factorization problem:

$$\min_{\mathbf{D},\mathbf{A}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 = \min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2$$

    where $\mathbf{A}$ is **sparse**, and $\mathbf{D}$ is the **dictionary**
  - Each patch is decomposed into $\mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i$
  - Average the reconstruction $\mathbf{D}\boldsymbol{\alpha}_i$ of each patch $\mathbf{x}_i$ to reconstruct a full-sized image

- The number of patches n is large ($=$ number of pixels)

Online optimization for dictionary learning

$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \ \text{ s.t. } \ \forall j = 1,\ldots,k, \ \ ||\mathbf{d}_j||_2 \leqslant 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$
- Good results, but very slow !

Online optimization for dictionary learning

$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

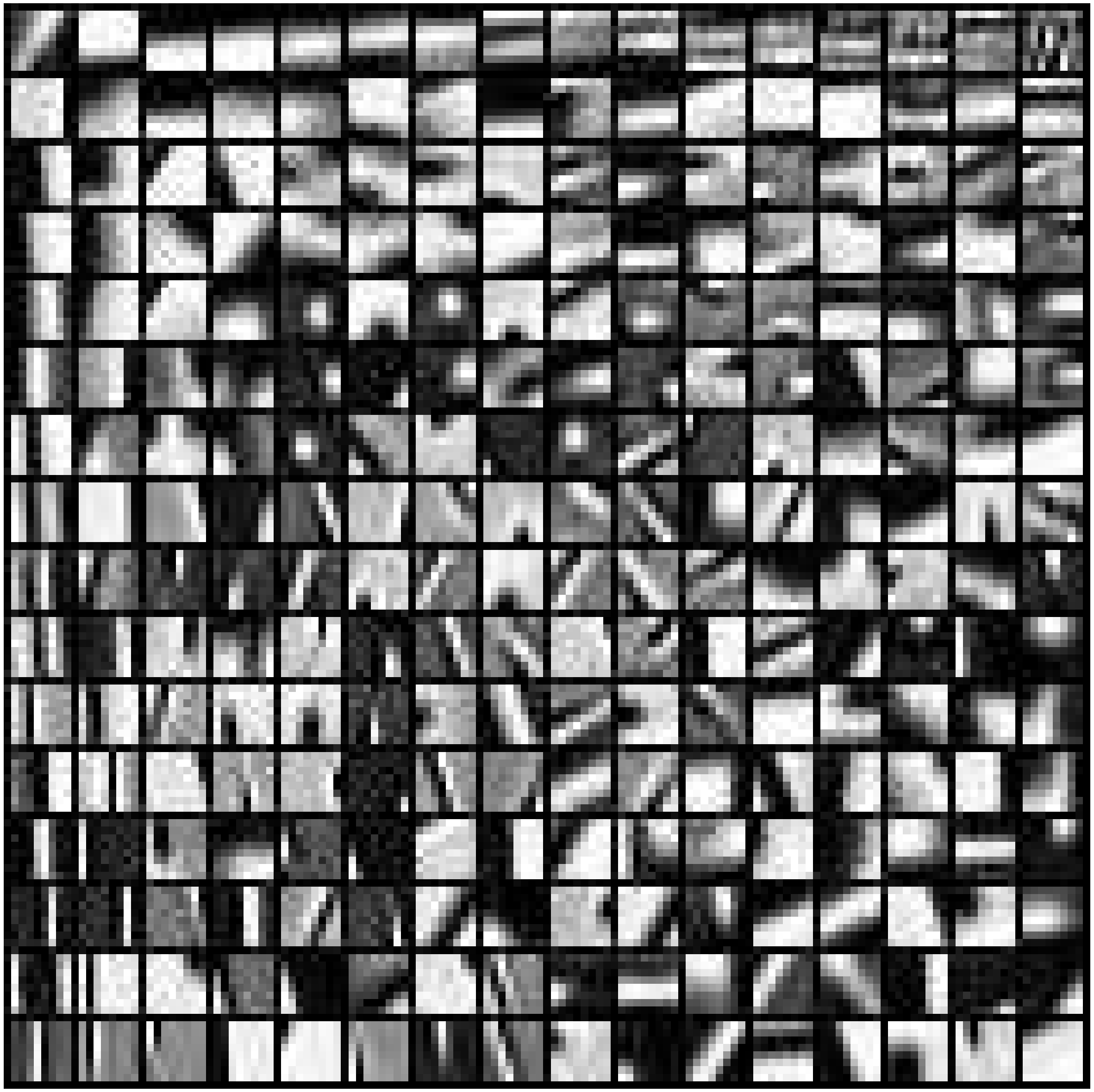
$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \text{ s.t. } \forall j = 1,\ldots,k, \quad ||\mathbf{d}_j||_2 \leqslant 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$.
- Good results, but very slow !
- **Online learning** (Mairal, Bach, Ponce, and Sapiro, 2009b) can
  - handle potentially infinite datasets
  - adapt to dynamic training sets
- **Simultaneous sparse coding** (Mairal et al., 2009e)
  - Links with NL-means (Buades et al., 2008)

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009e)

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009e)

What does the dictionary $\mathbf{D}$ look like?



Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

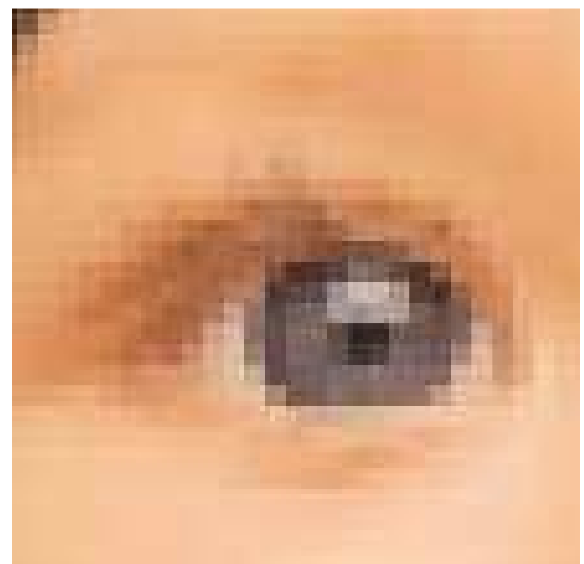
Additional methods - Softwares

- 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/`

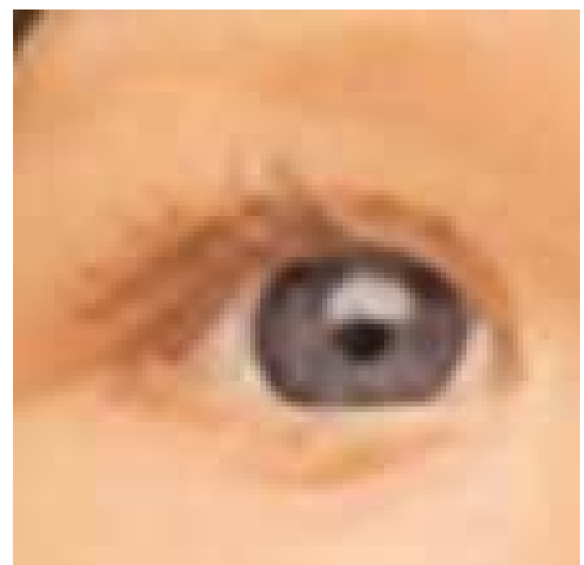
Task-driven dictionary learning (Mairal, Bach, and Ponce, 2010a)

- Define $\alpha^*(D, x) = \operatorname{argmin}_\alpha \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1$
- α is used as a code for x
- **Direct optimization of $\alpha^*(D, x)$ with respect to D**
  - Application to image processing tasks such inverse half-toning (Mairal, Bach, and Ponce, 2010a)
  - Image super-resolution (Couzinie-Devy, Mairal, Bach, and Ponce, 2011)

Digital Zooming (Couzinie-Devy et al., 2011)



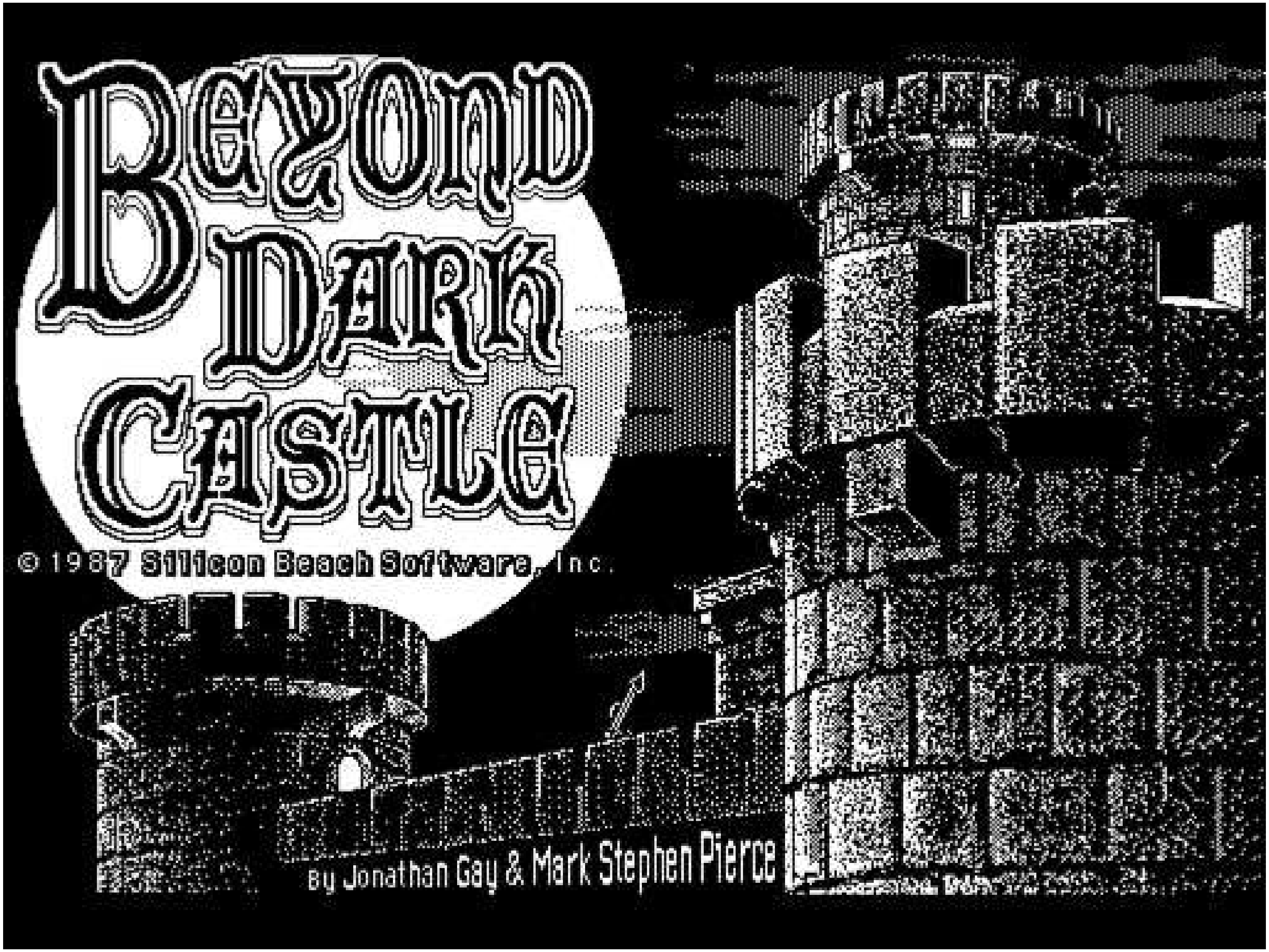
Digital Zooming (Couzinie-Devy et al., 2011)



Inverse half-toning (Mairal et al., 2011)

Inverse half-toning (Mairal et al., 2011)

Ongoing Work - Inverse half-toning



Ongoing Work - Inverse half-toning

Ongoing Work - Inverse half-toning

Ongoing Work - Inverse half-toning

Outline

- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices
- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

Sparsity in supervised machine learning

- Observed data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
  - Response vector $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$
  - Design matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$

- Regularized empirical risk minimization:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w) = \boxed{\min_{w \in \mathbb{R}^p} \; L(y, Xw) + \lambda \Omega(w)}$$

- Norm Ω to promote sparsity
  - square loss + ℓ_1-norm $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$

Sparsity in unsupervised machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1,\dots,w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}$$

Sparsity in unsupervised machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1,\dots,w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \Big\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \Big\}$$

- **Only responses are observed** $\Rightarrow$ **Dictionary learning**
  - Learn $X = (x^1, \dots, x^p) \in \mathbb{R}^{n \times p}$ such that $\forall j, \ \|x^j\|_2 \leqslant 1$

$$\min_{X=(x^1,\dots,x^p)} \min_{w^1,\dots,w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \Big\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \Big\}$$

  - Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

- **sparse PCA**: replace $\|x^j\|_2 \leqslant 1$ by $\Theta(x^j) \leqslant 1$

Sparsity in signal processing

- **Multiple** responses/signals $x = (x^1, \dots, x^k) \in \mathbb{R}^{n \times k}$

$$\min_{\alpha^1,\dots,\alpha^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

- **Only responses are observed** $\Rightarrow$ **Dictionary learning**
  - Learn $D = (d^1, \dots, d^p) \in \mathbb{R}^{n \times p}$ such that $\forall j, \ \|d^j\|_2 \leqslant 1$

$$\min_{D=(d^1,\dots,d^p)} \min_{\alpha^1,\dots,\alpha^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

  - Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

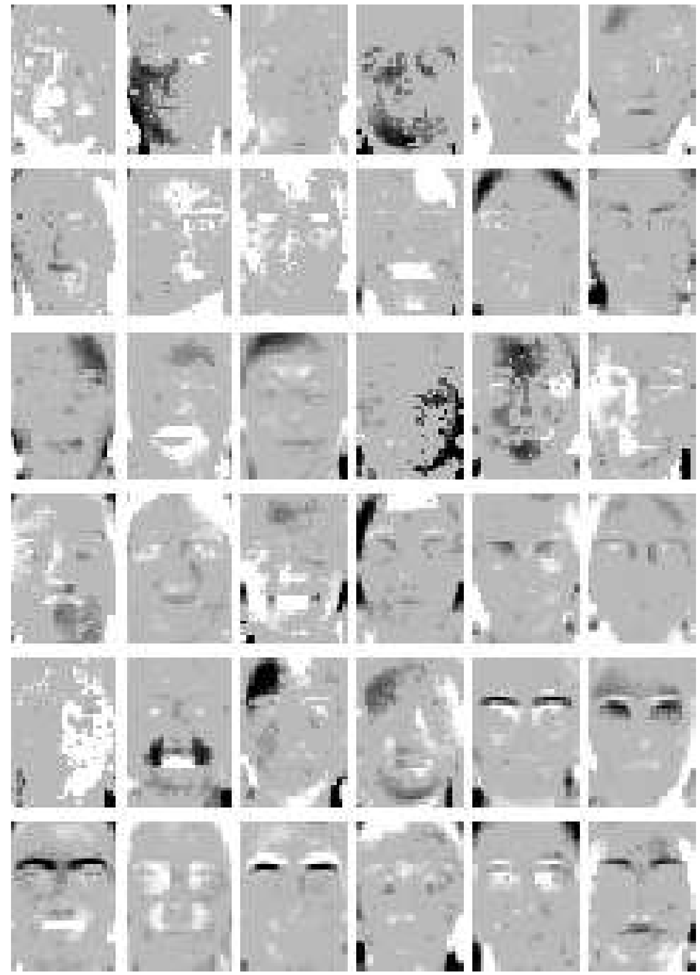
- **sparse PCA**: replace $\|d^j\|_2 \leqslant 1$ by $\Theta(d^j) \leqslant 1$

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010b)

Structured sparse PCA (Jenatton et al., 2009b)

raw data

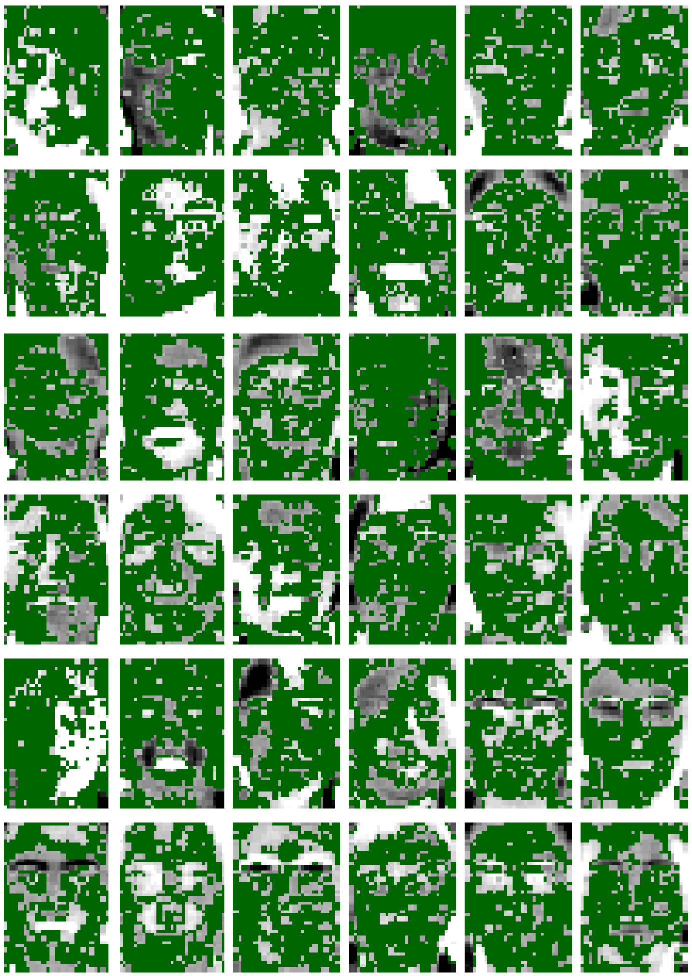


sparse PCA

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)

raw data

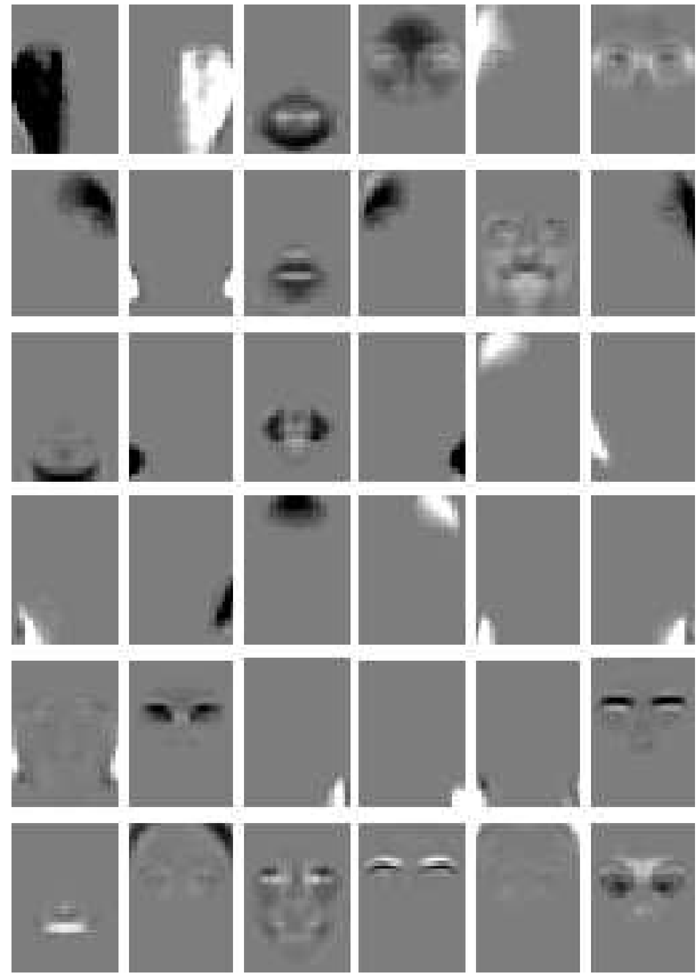


sparse PCA

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)

raw data

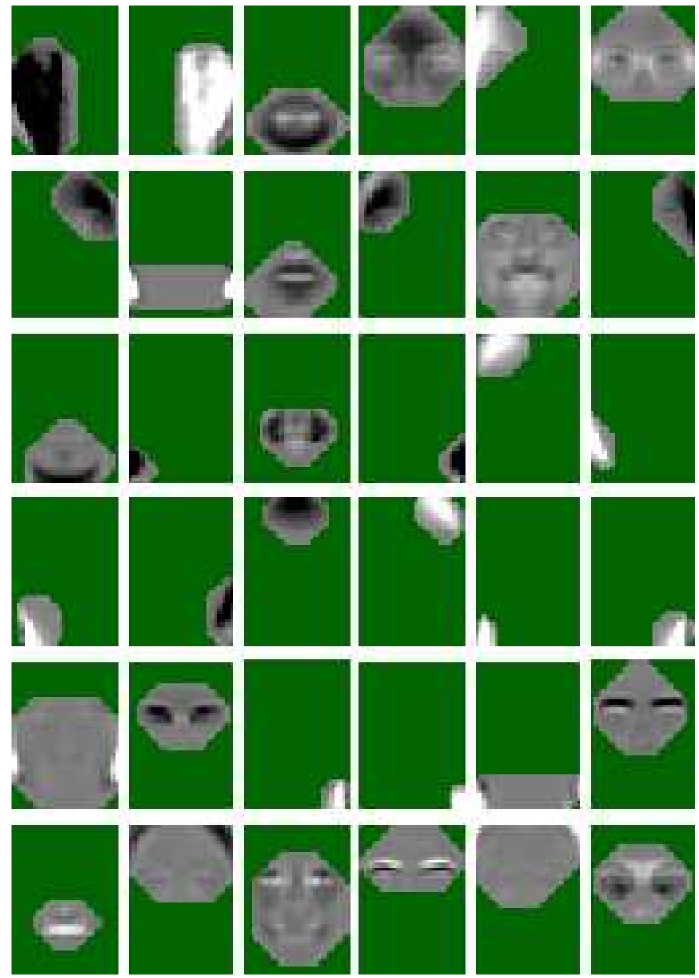


Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion in face identification

Structured sparse PCA (Jenatton et al., 2009b)

raw data



Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion in face identification

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements “organized” in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010b)

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010b)
- **Stability and identifiability**
  - Optimization problem $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \|w\|_1$ is unstable
  - "Codes" w^j often used in later processing (Mairal et al., 2009d)
- **Prediction or estimation performance**
  - When prior knowledge matches data (Haupt and Nowak, 2006; Baraniuk et al., 2008; Jenatton et al., 2009a; Huang et al., 2009)
- **Numerical efficiency**
  - Non-linear variable selection with 2^p subsets (Bach, 2008c)

Classical approaches to structured sparsity

- **Many application domains**
  - Computer vision (Cevher et al., 2008; Mairal et al., 2009c)
  - Neuro-imaging (Gramfort and Kowalski, 2009; Jenatton et al., 2011)
  - Bio-informatics (Rapaport et al., 2008; Kim and Xing, 2010)

- **Non-convex approaches**
  - Haupt and Nowak (2006); Baraniuk et al. (2008); Huang et al. (2009)

- **Convex approaches**
  - Design of sparsity-inducing norms

Outline

- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices

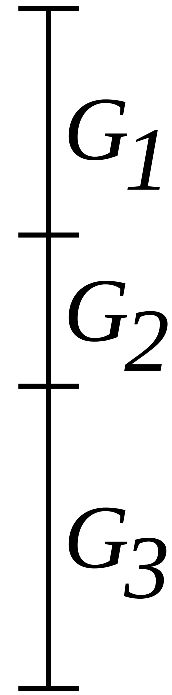
- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications

- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

Sparsity-inducing norms

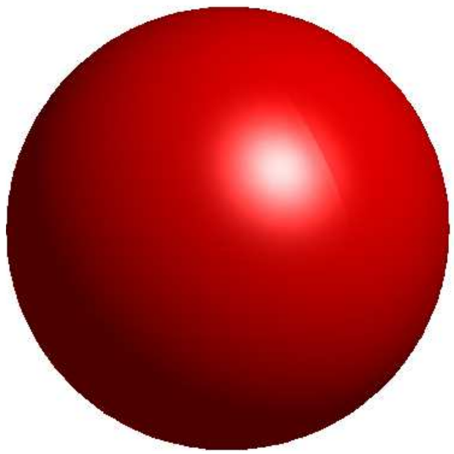
- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$

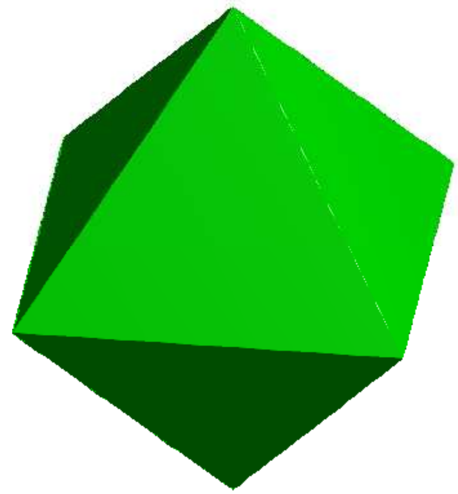


  - with $\mathbf{H}$ a partition of $\{1, \dots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

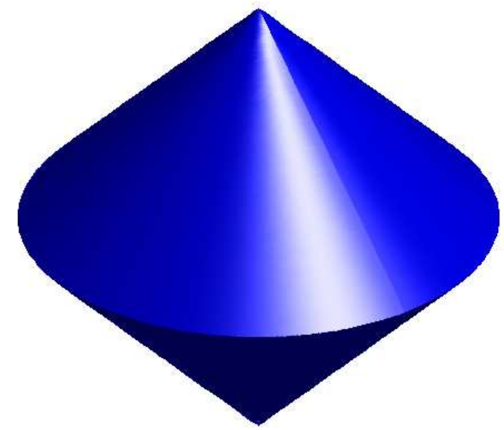
Unit norm balls
Geometric interpretation



$\|w\|_2$



$\|w\|_1$

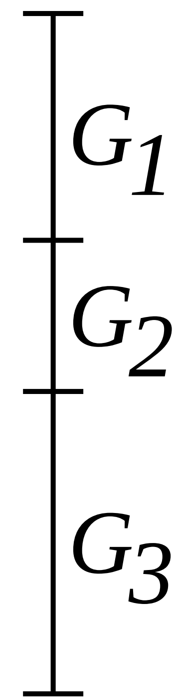


$\sqrt{w_1^2 + w_2^2} + |w_3|$

Sparsity-inducing norms

- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$

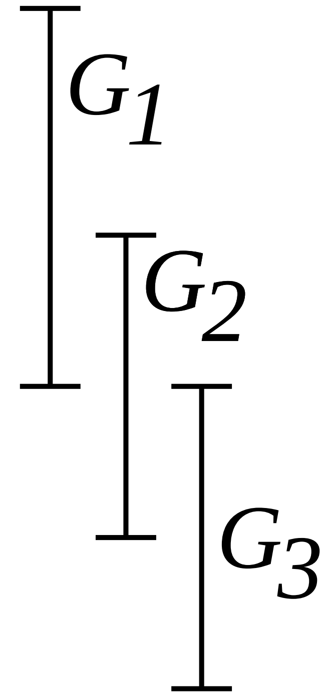


  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

- However, the ℓ_1-ℓ_2 norm encodes **fixed/static prior information**, requires to know in advance how to group the variables

- What happens if the set of groups $\mathbf{H}$ is not a partition anymore?

Structured sparsity with overlapping groups (Jenatton, Audibert, and Bach, 2009a)

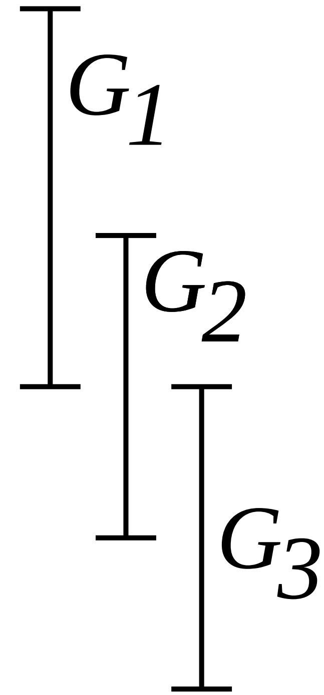


- When penalizing by the ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$

  - The ℓ_1 norm induces sparsity at the group level:
    - Some w_G's are set to zero
  - Inside the groups, the ℓ_2 norm does not promote sparsity

Structured sparsity with overlapping groups (Jenatton, Audibert, and Bach, 2009a)



- When penalizing by the ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$

  – The ℓ_1 norm induces sparsity at the group level:
    * Some w_G's are set to zero
  – Inside the groups, the ℓ_2 norm does not promote sparsity

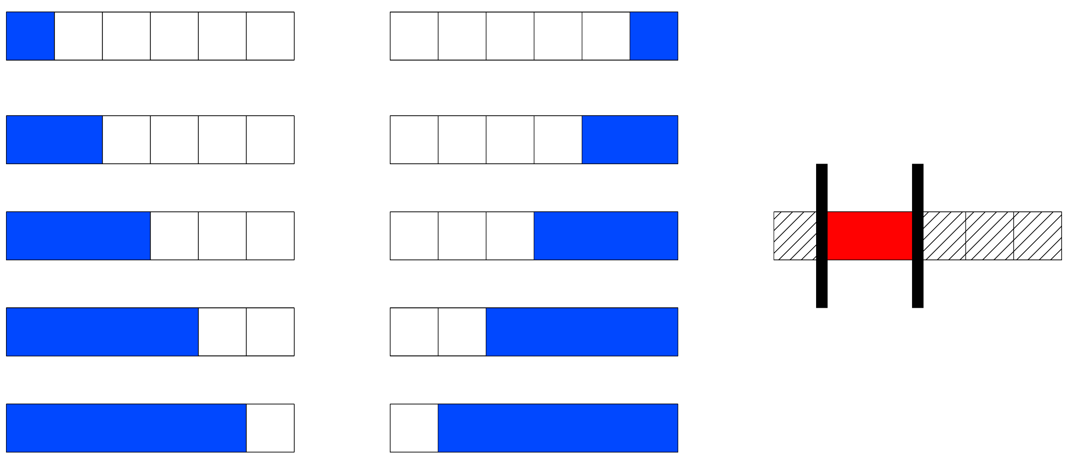
- The zero pattern of w is given by

$$\{j, \; w_j = 0\} = \bigcup_{G \in \mathbf{H}'} G \;\; \text{for some } \mathbf{H}' \subseteq \mathbf{H}$$

- **Zero patterns are unions of groups**

Examples of set of groups $\mathbf{H}$

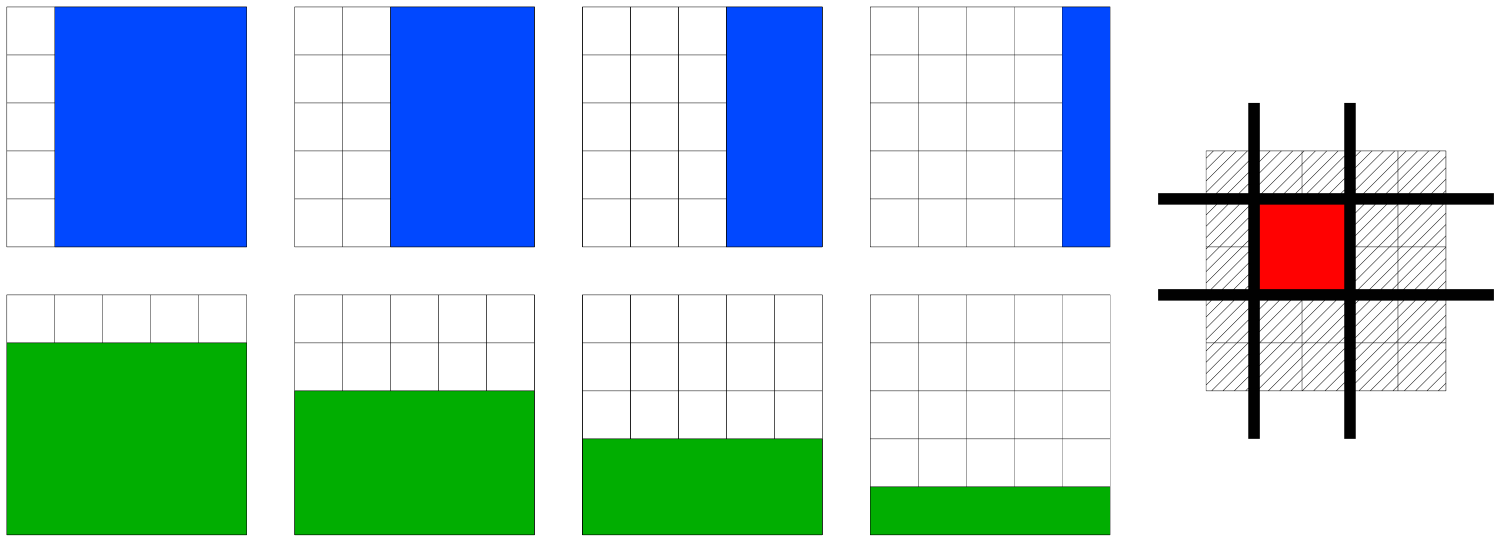
- Selection of contiguous patterns on a sequence, $p = 6$



  - $\mathbf{H}$ is the set of blue groups
  - Any union of blue groups set to zero leads to the selection of a contiguous pattern

Examples of set of groups $\mathbb{H}$

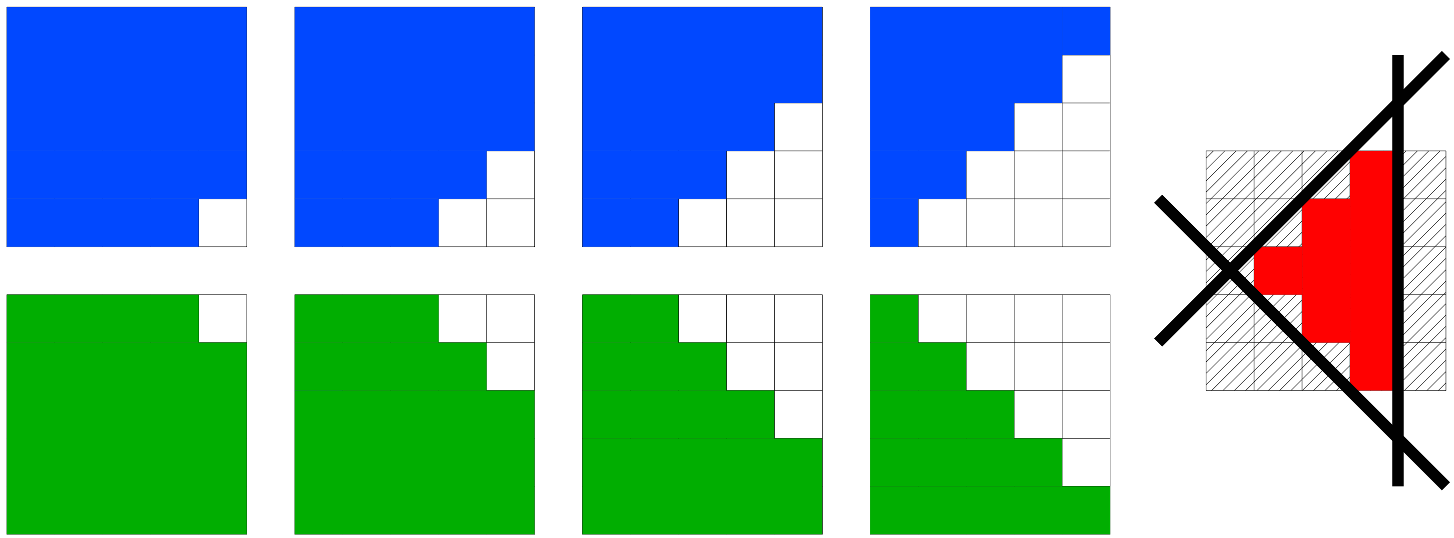
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ is the set of blue/green groups (with their not displayed complements)
  - Any union of blue/green groups set to zero leads to the selection of a rectangle

Examples of set of groups $\mathbb{H}$

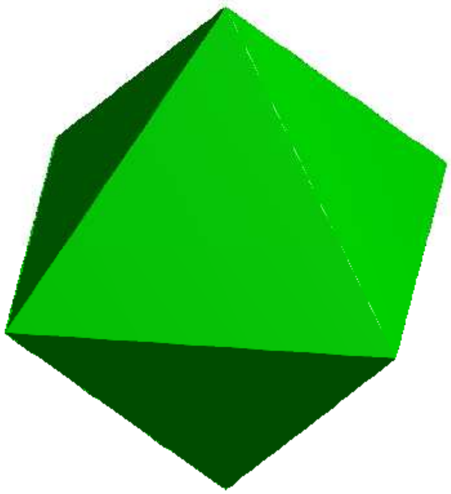
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



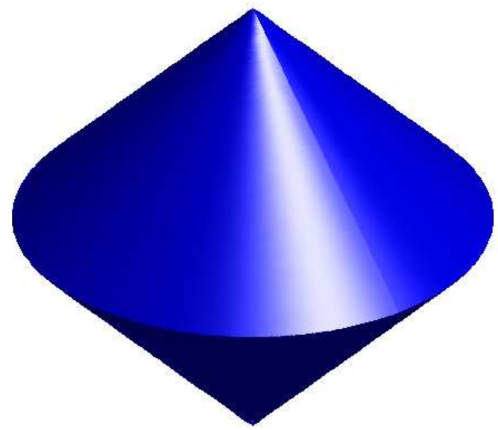
  - It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

Geometric interpretation



$\|w\|_1$



$\sqrt{w_1^2 + w_2^2} + |w_3|$

$\|w\|_2 + |w_1| + |w_2|$

Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2011)

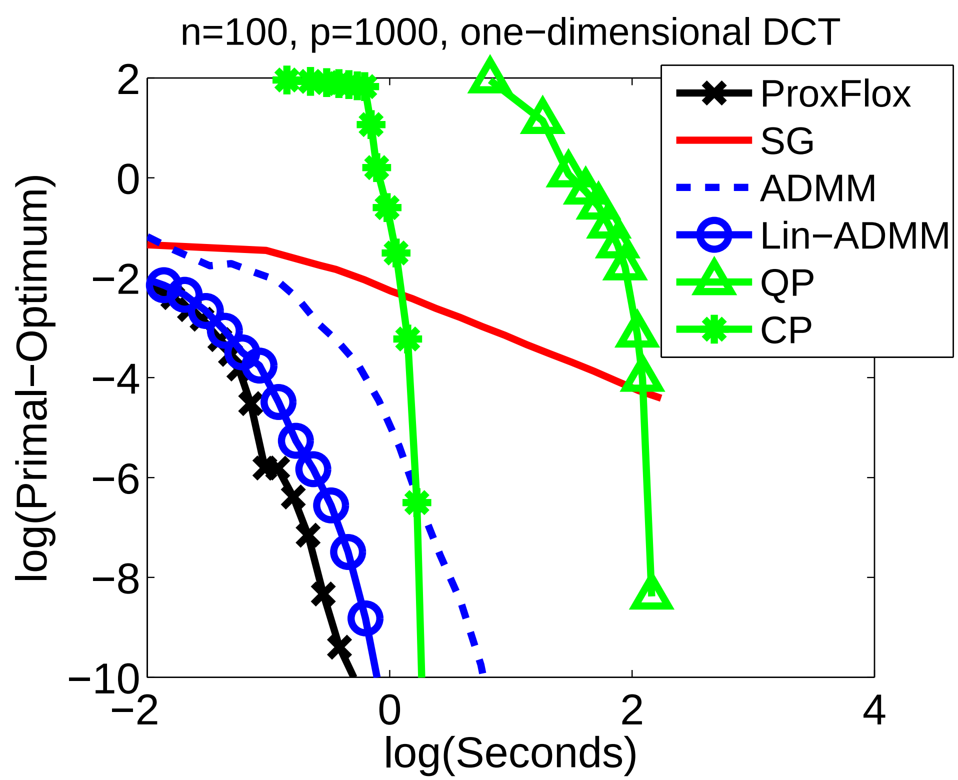
- Gradient descent as a **proximal method** (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2} \|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{B} \nabla L(w_t)$

Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{B}\nabla L(w_t)$
- Problems of the form: $\boxed{\min_{w \in \mathbb{R}^p} L(w) + \lambda\Omega(w)}$
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda\Omega(w) + \frac{B}{2}\|w - w_t\|_2^2$
  - $\Omega(w) = \|w\|_1 \Rightarrow$ **Thresholded gradient descent**
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

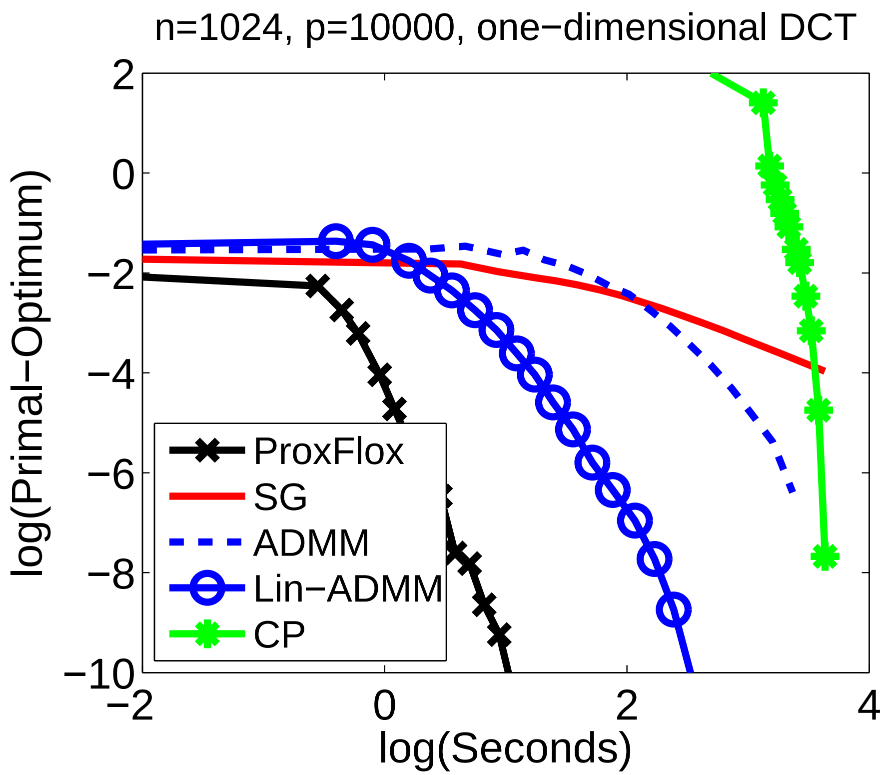
Comparison of optimization algorithms
(Mairal, Jenatton, Obozinski, and Bach, 2010b)
Small scale

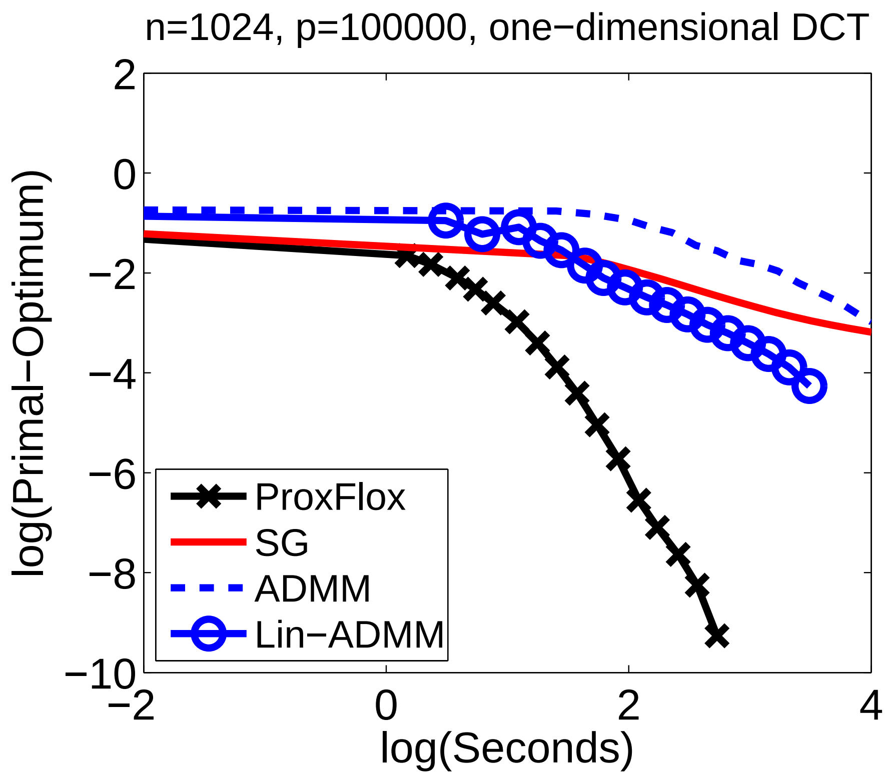
- Specific norms which can be implemented through network flows



Comparison of optimization algorithms
(Mairal, Jenatton, Obozinski, and Bach, 2010b)
Large scale

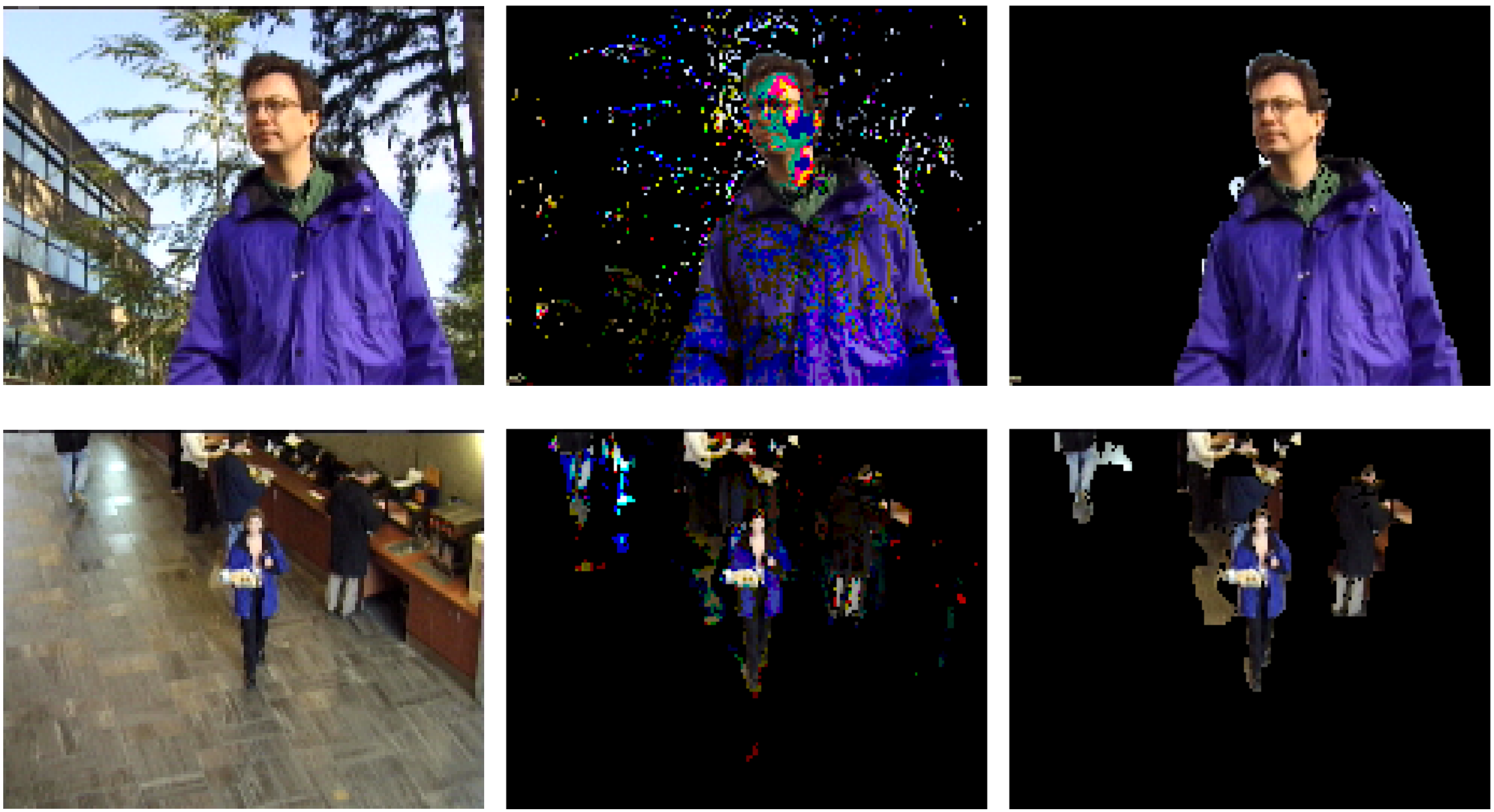
- Specific norms which can be implemented through network flows





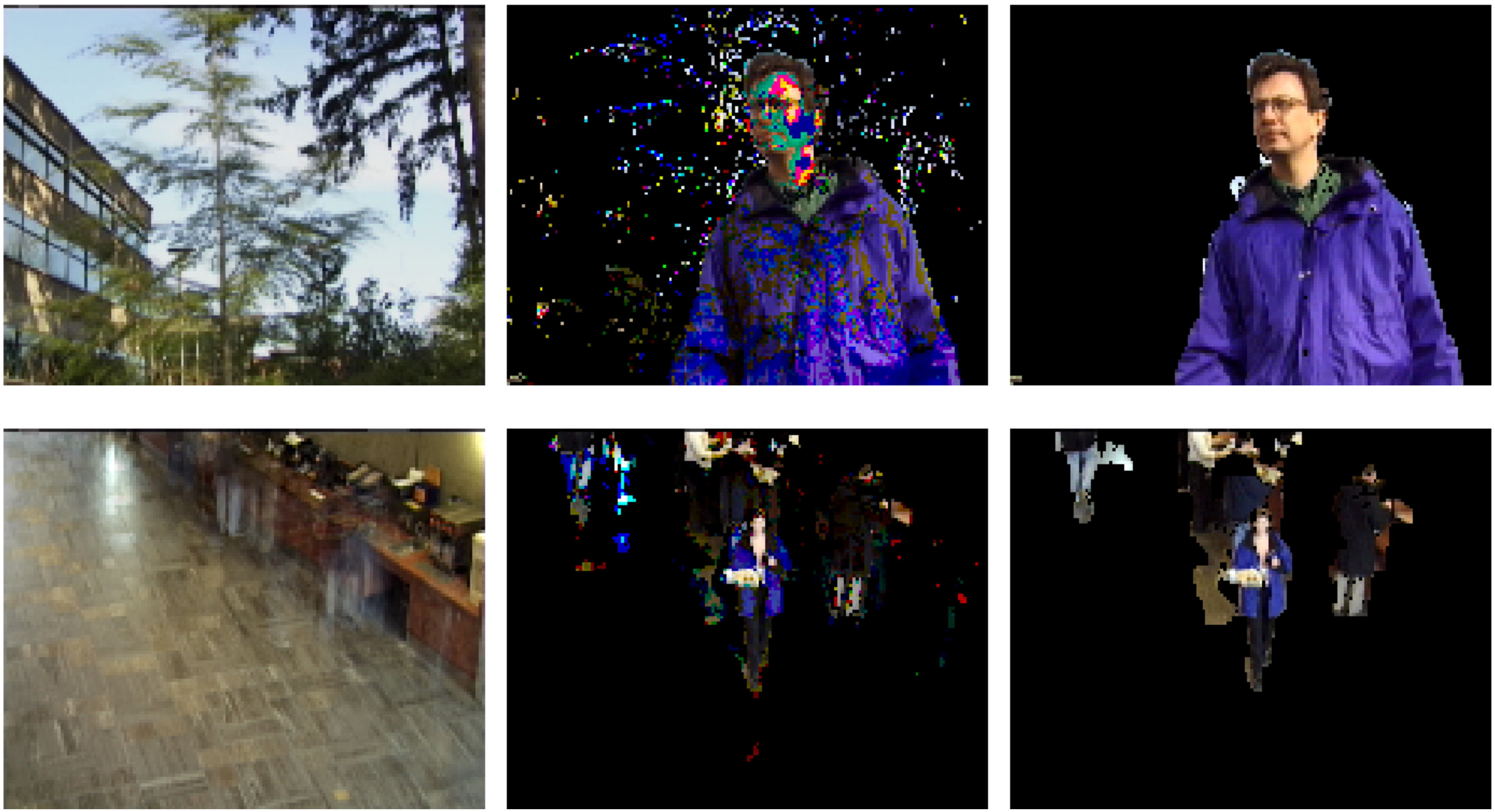
Application to background subtraction
(Mairal, Jenatton, Obozinski, and Bach, 2010b)

Input | ℓ_1-norm | Structured norm



Application to background subtraction
(Mairal, Jenatton, Obozinski, and Bach, 2010b)

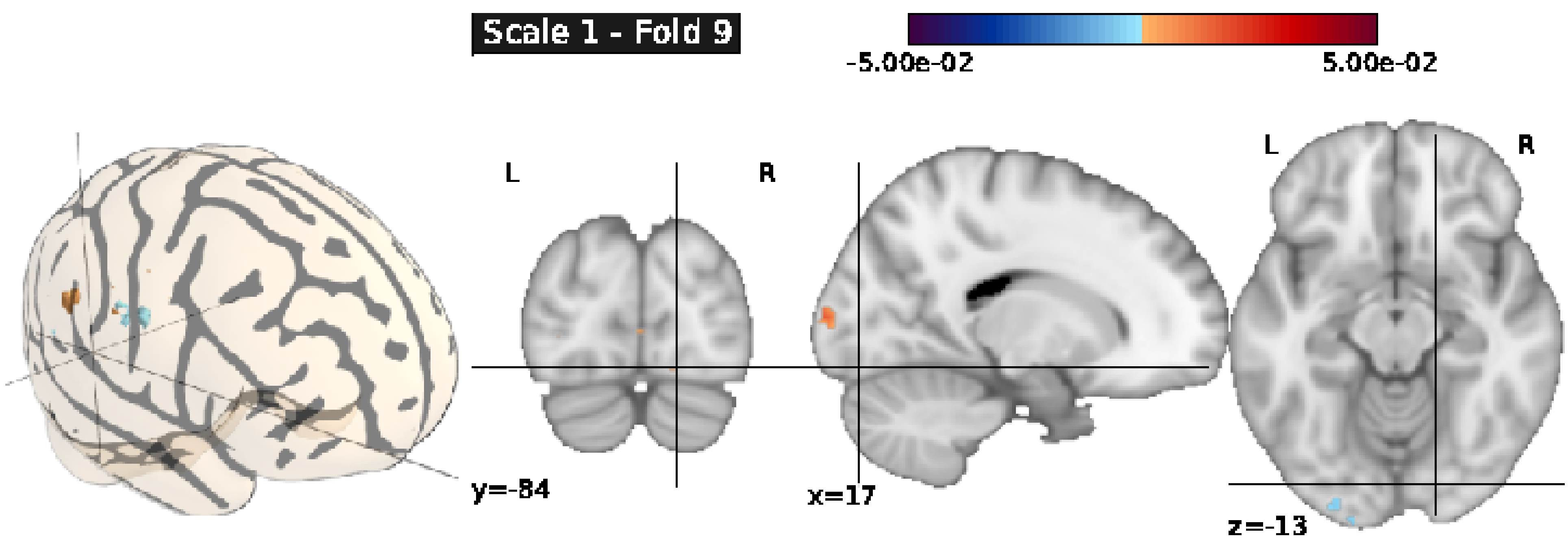
Background | ℓ_1-norm | Structured norm



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

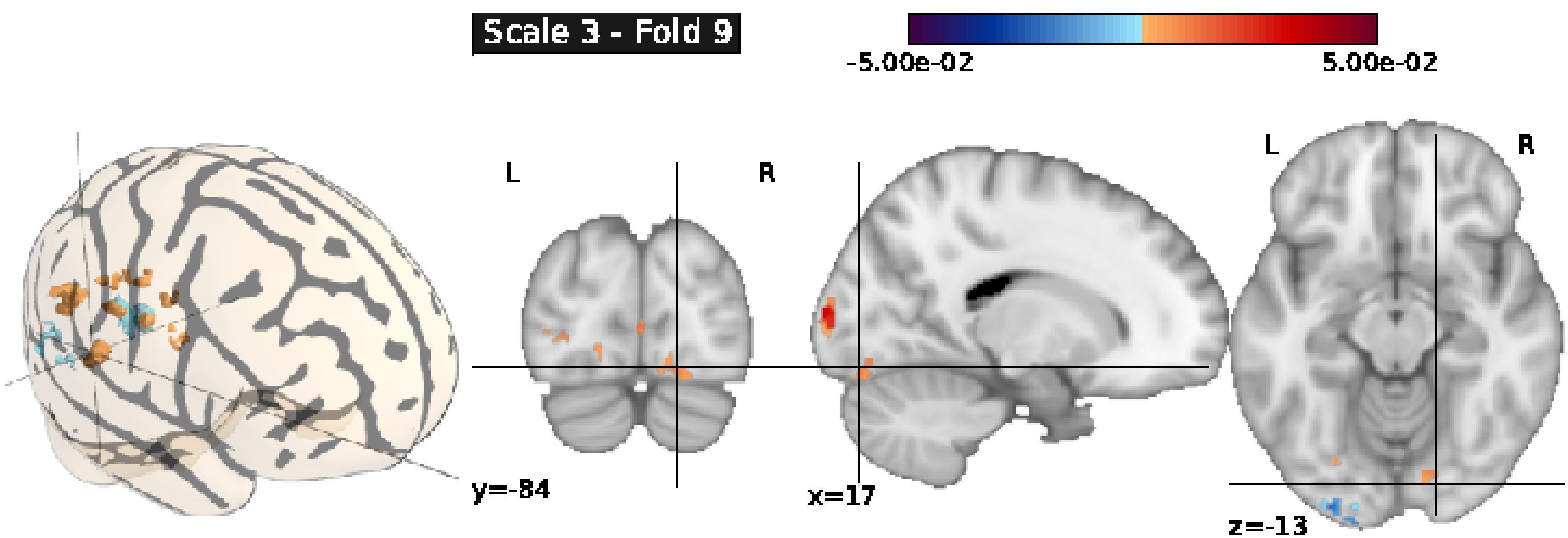
- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

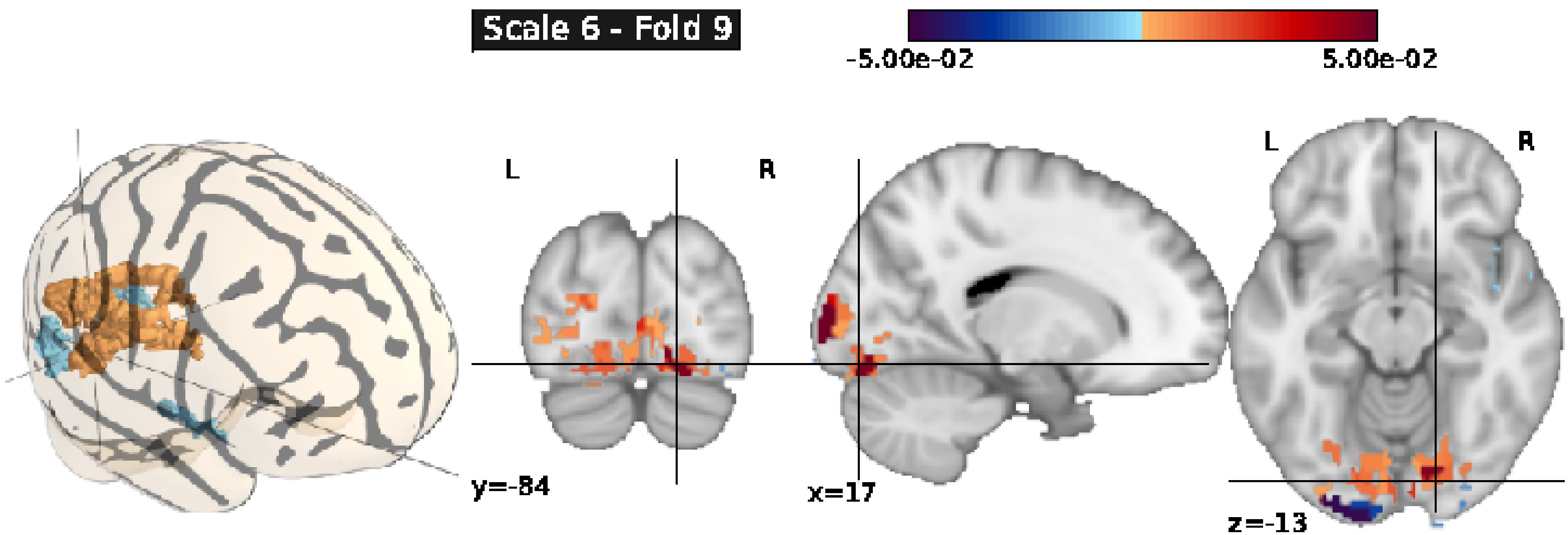
Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging
Structured sparsity for fMRI (Jenatton et al., 2011)

- "Brain reading": prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization

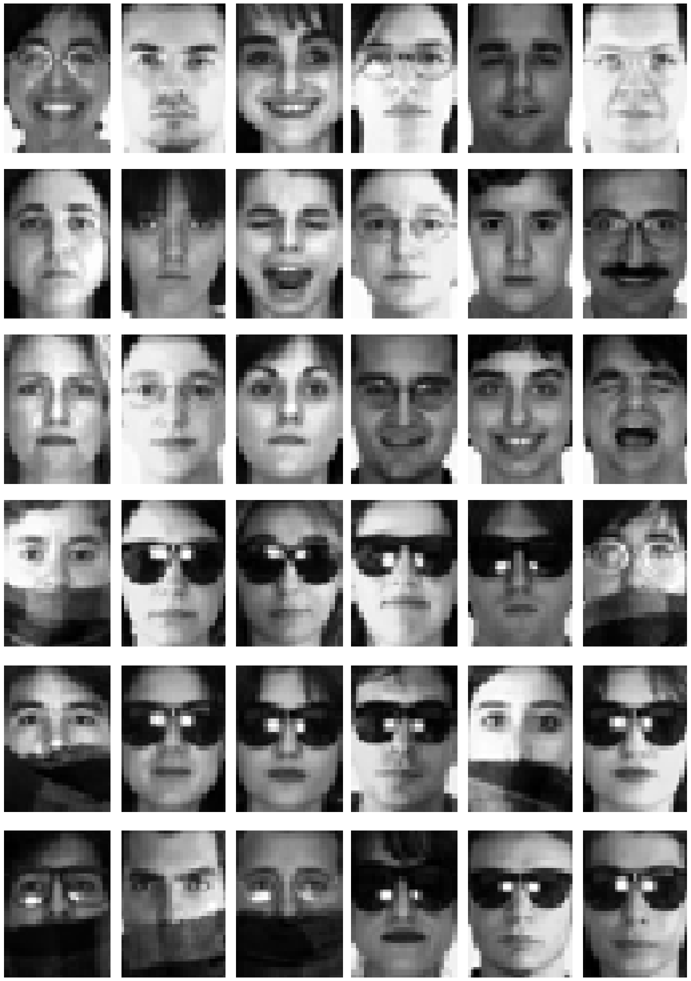


Sparse Structured PCA
(Jenatton, Obozinski, and Bach, 2009b)

- Learning **sparse and structured dictionary elements**:

$$\min_{W\in\mathbb{R}^{k\times n}, X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i - Xw^i\|_2^2 + \lambda\sum_{j=1}^{p}\Omega(x^j) \text{ s.t. } \forall i,\ \|w^i\|_2 \leq 1$$

Application to face databases (1/3)

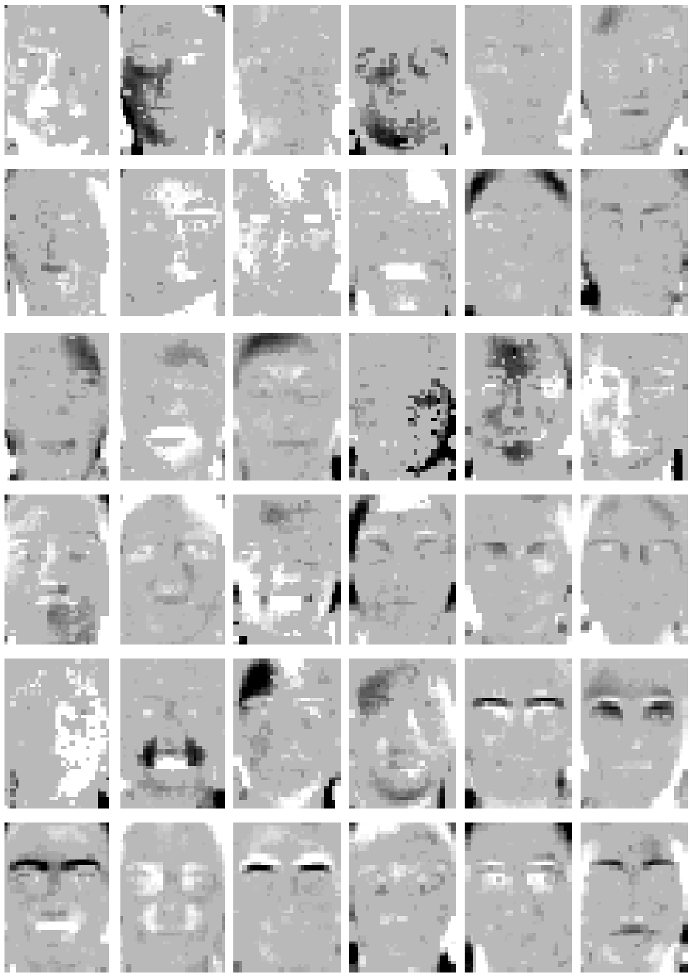


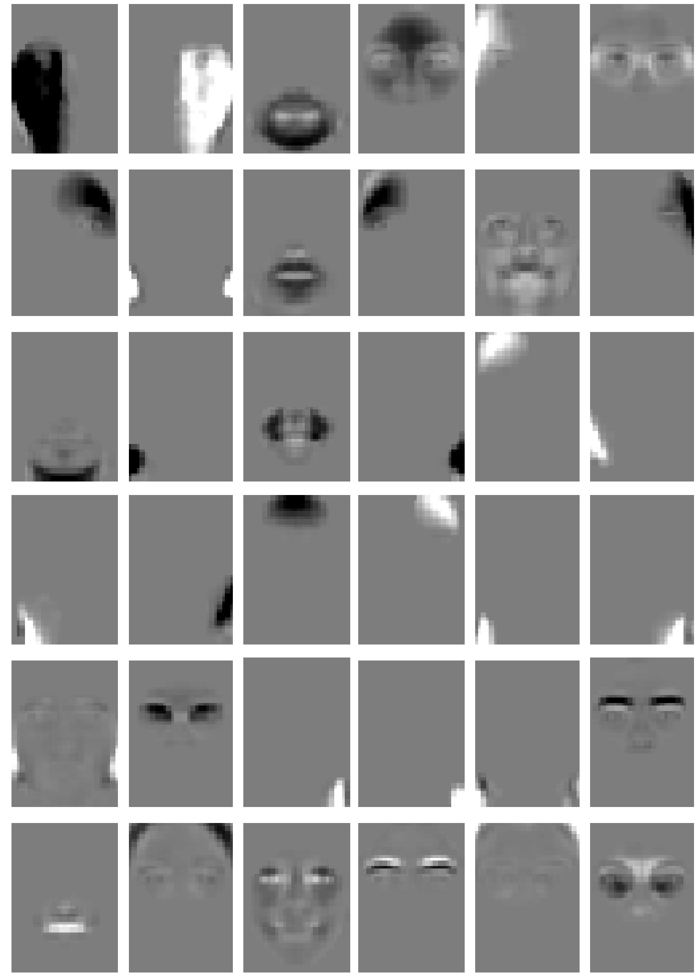
raw data

(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)

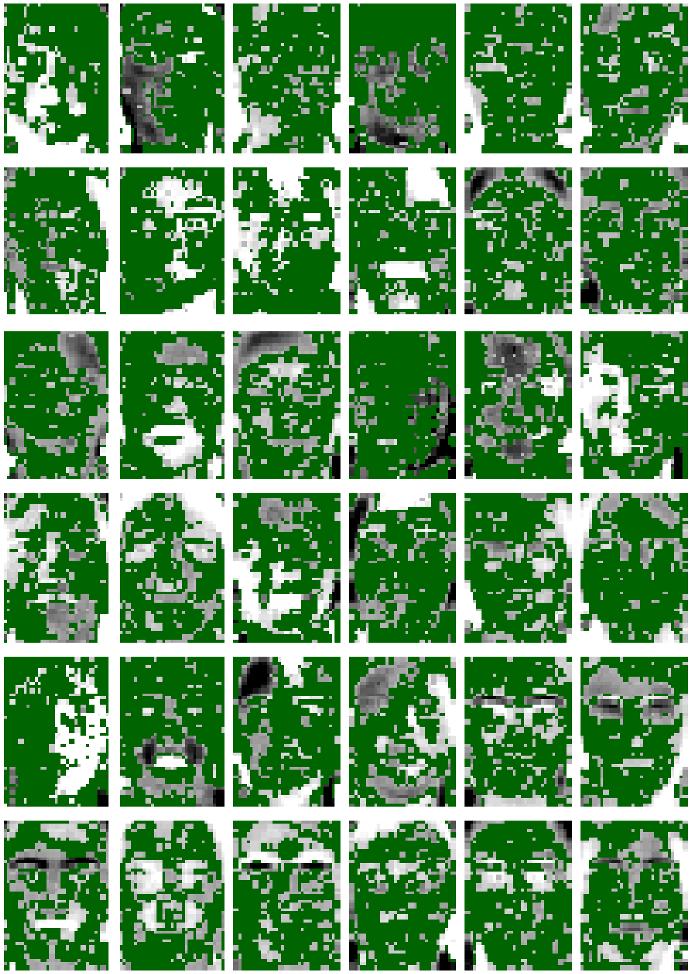




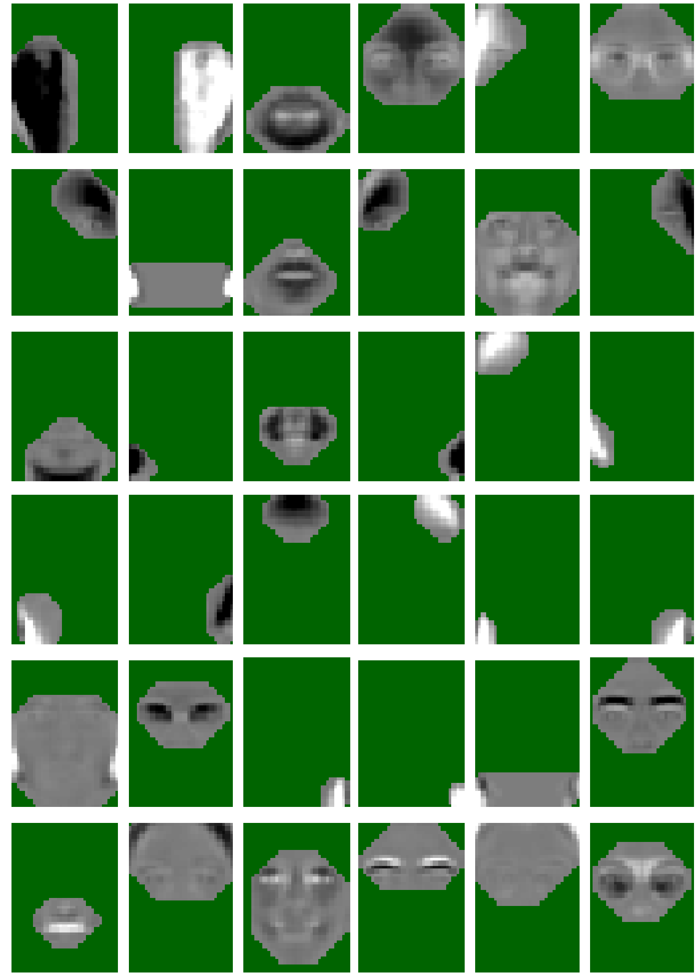
(unstructured) sparse PCA   Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

Application to face databases (2/3)



(unstructured) sparse PCA

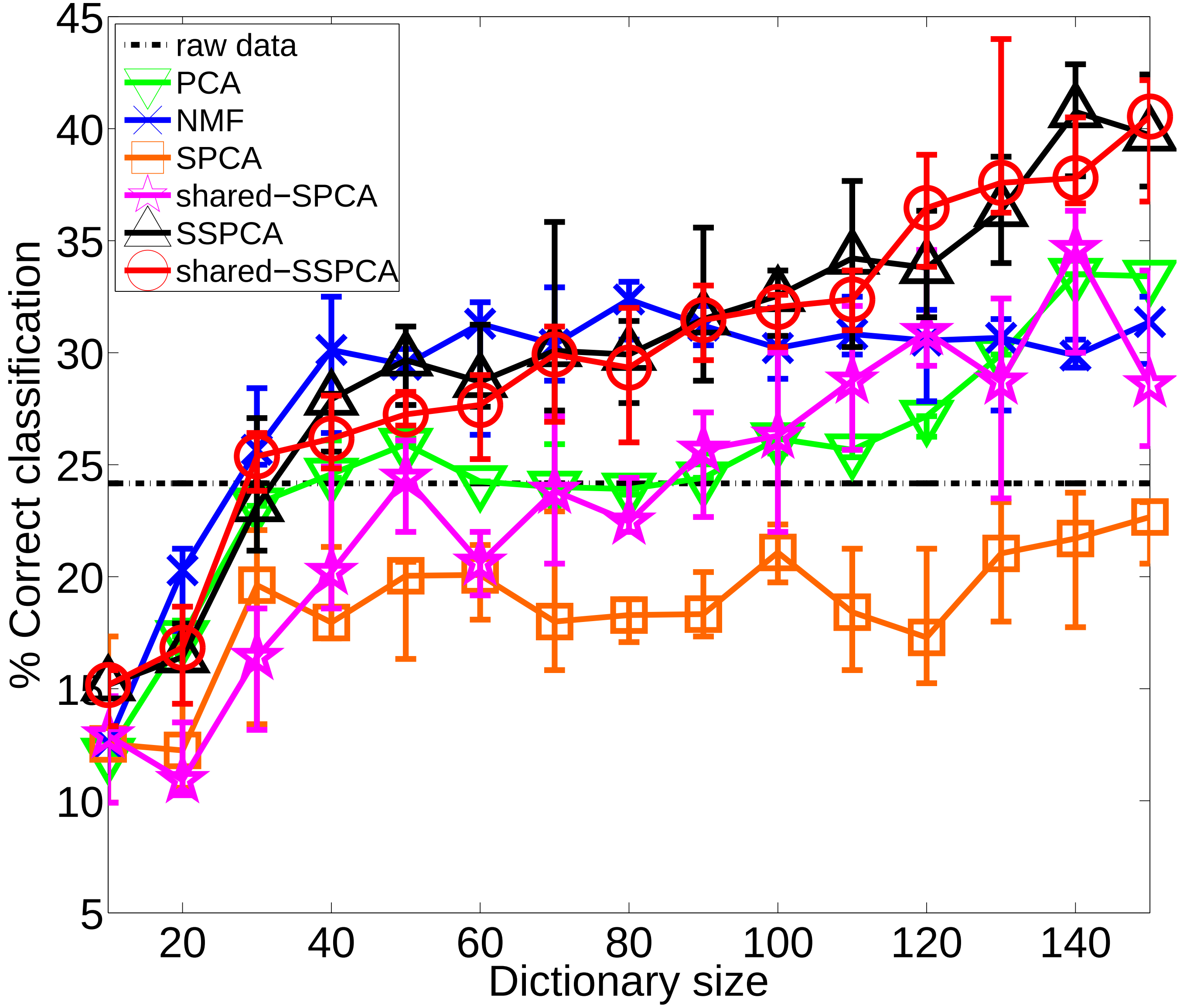


Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

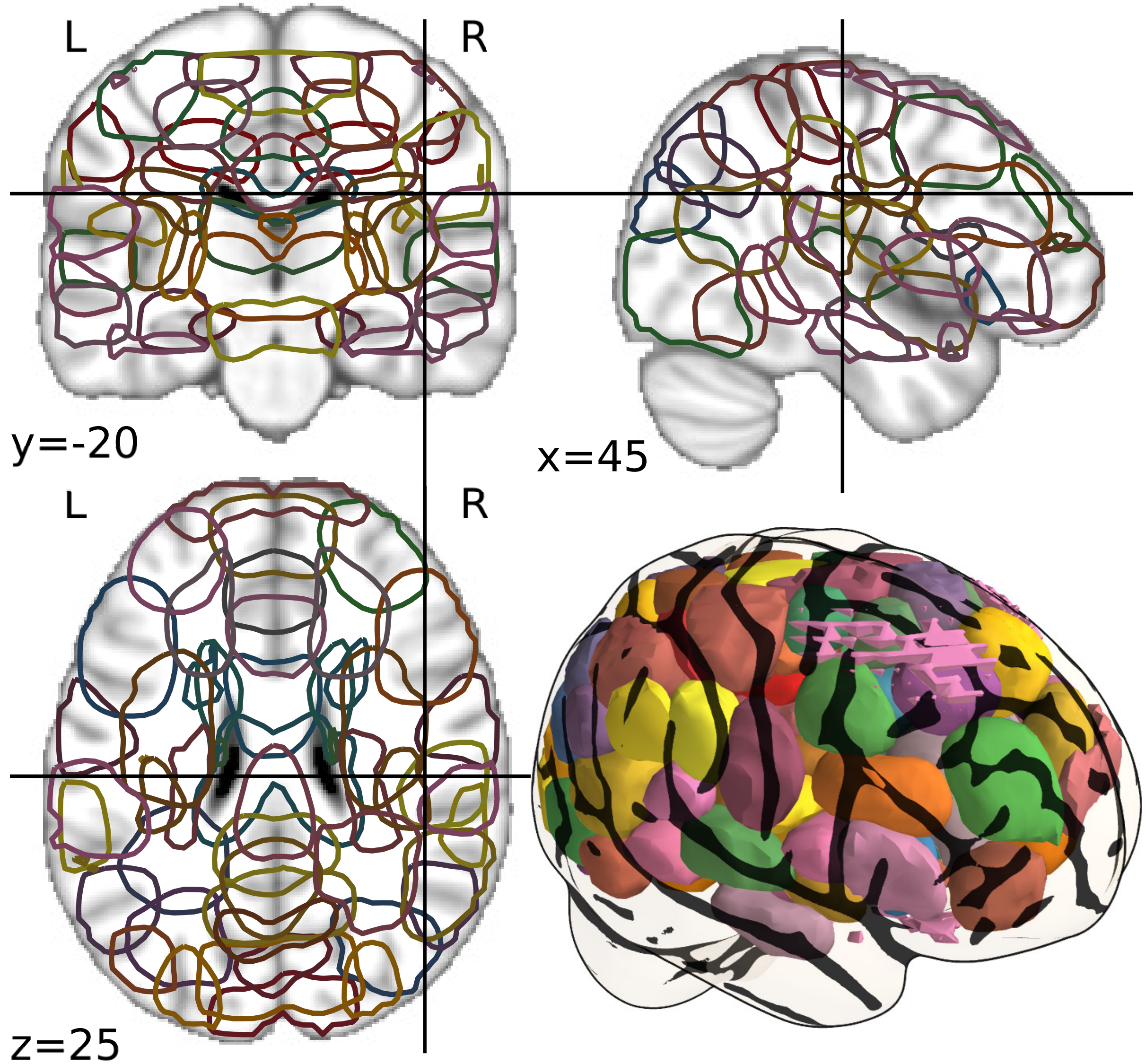
Application to face databases (3/3)

- Quantitative performance evaluation on classification task



Structured sparse PCA on resting state activity

(Varoquaux, Jenatton, Gramfort, Obozinski, Thirion, and Bach, 2010)



Dictionary learning vs. sparse structured PCA
Exchange roles of X and w

- Sparse structured PCA (**structured dictionary elements**):

$$\min_{W\in\mathbb{R}^{k\times n},X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i-Xw^i\|_2^2+\lambda\sum_{j=1}^{k}\Omega(x^j) \text{ s.t. } \forall i,\ \|w^i\|_2\leq 1.$$

- Dictionary learning with **structured sparsity for codes** w:

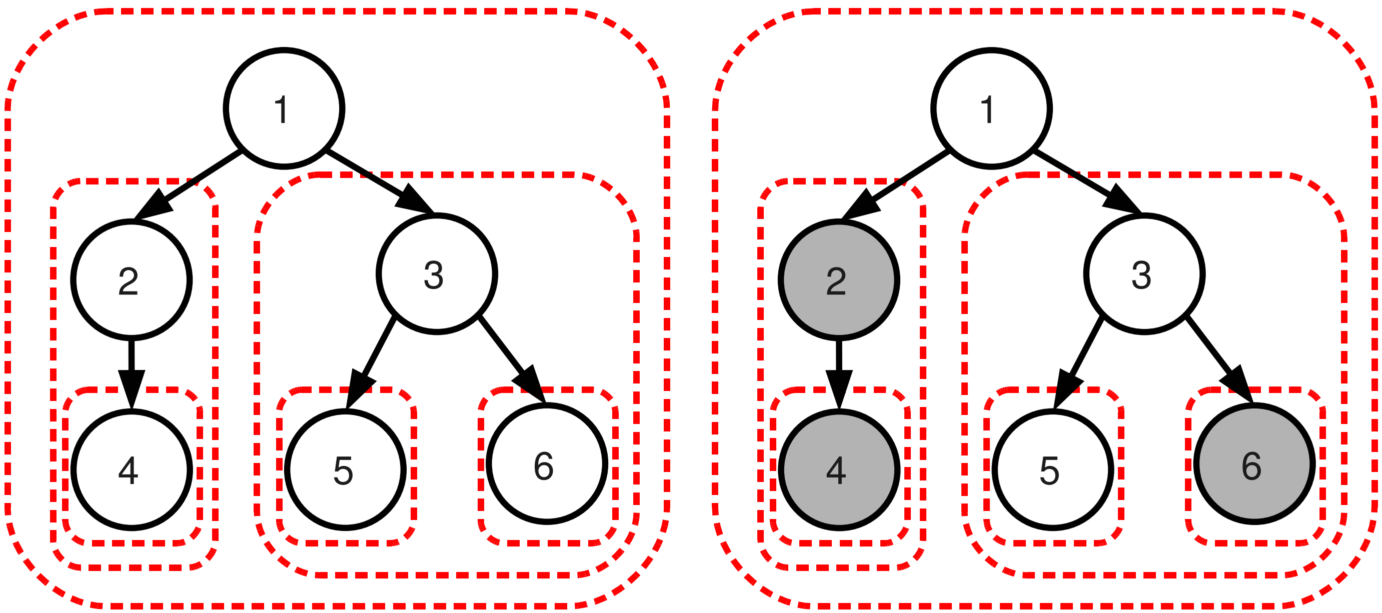
$$\min_{W\in\mathbb{R}^{k\times n},X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i-Xw^i\|_2^2+\lambda\Omega(w^i) \text{ s.t. } \forall j,\ \|x^j\|_2\leq 1.$$

- **Optimization**:
  - Alternating optimization
  - **Modularity of implementation** if proximal step is efficient (Jenatton et al., 2010; Mairal et al., 2010b)

Hierarchical dictionary learning (Jenatton, Mairal, Obozinski, and Bach, 2010)

- Structure on codes w (not on dictionary X)

- Hierarchical penalization: $\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_2$ where groups G in $\mathbf{H}$ are equal to set of descendants of some nodes in a tree

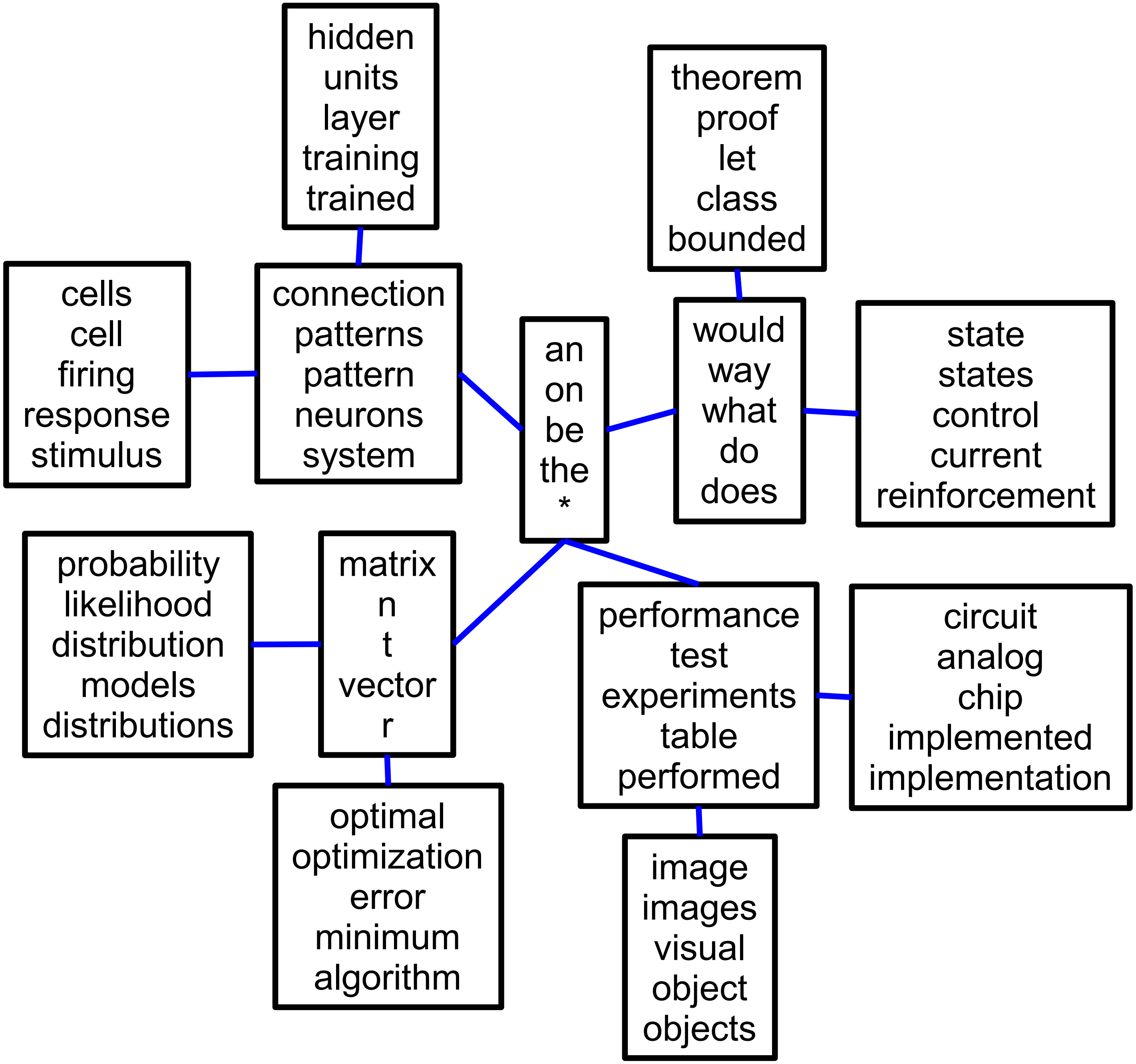


- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008c)

Hierarchical dictionary learning
Modelling of text corpora

- Each document is modelled through word counts
  - Low-rank matrix factorization of word-document matrix
  - Similar to NMF with multinomial loss
- Probabilistic topic models (Blei et al., 2003)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**

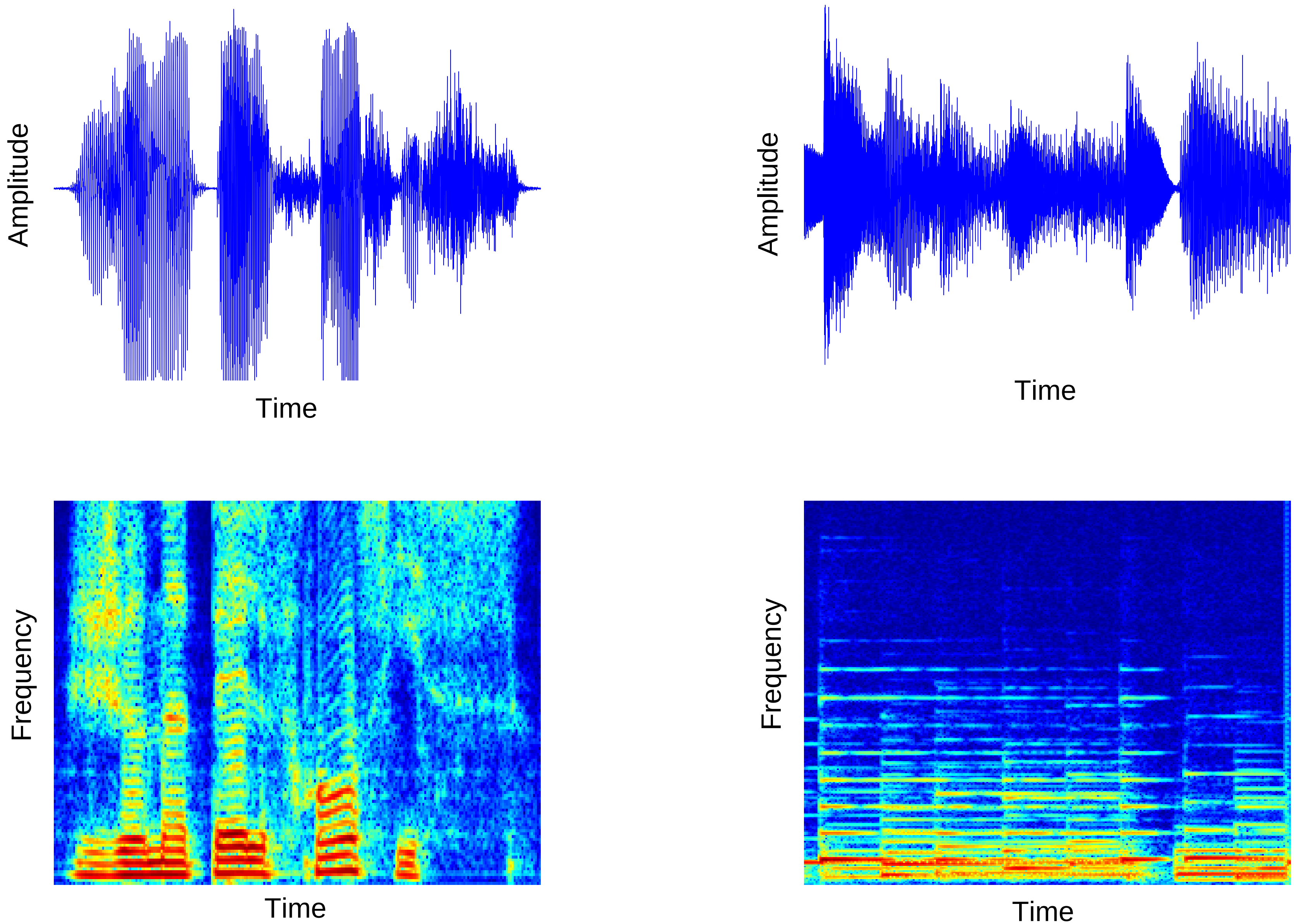
Modelling of text corpora - Dictionary tree



Topic models, NMF and matrix factorization

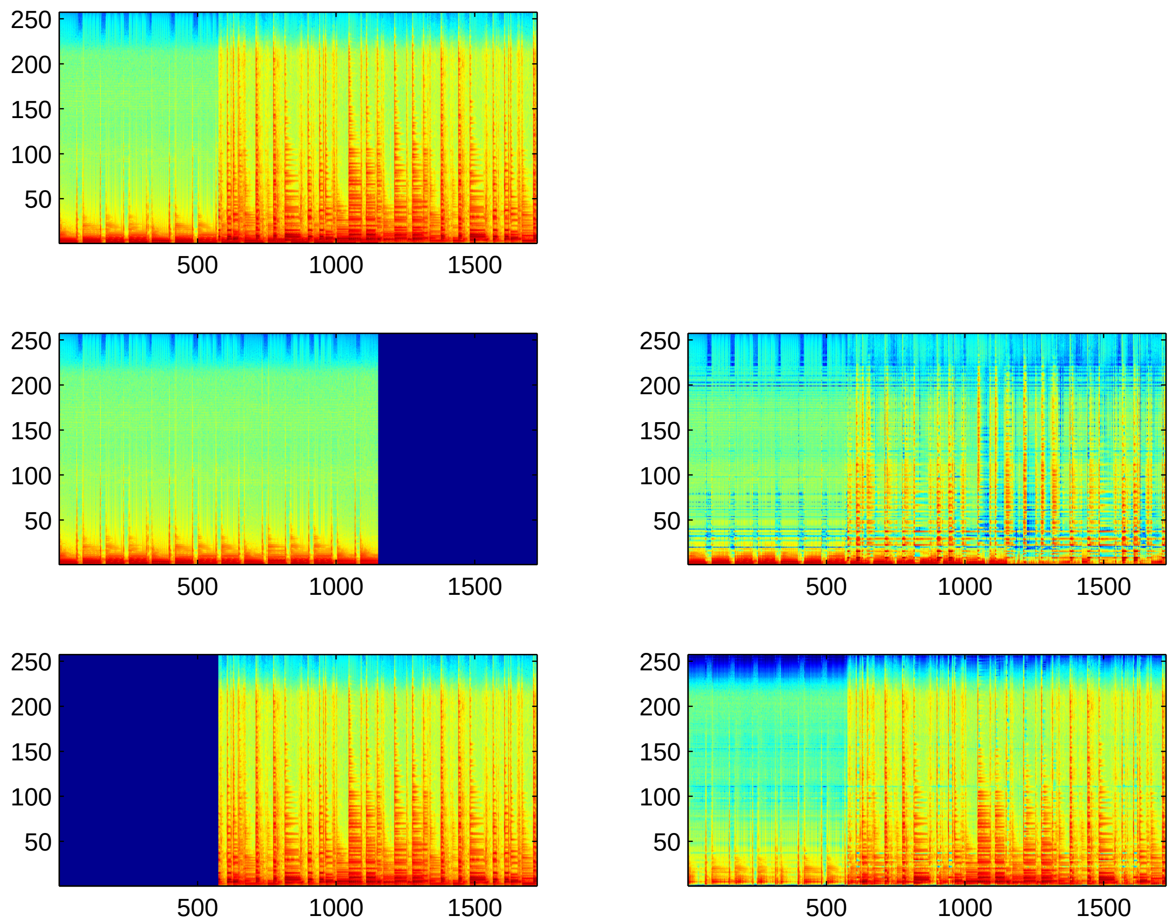
- **Three different views on the same problem**
  - Interesting parallels to be made
  - Common problems to be solved
- **Structure on dictionary/decomposition coefficients** with adapted priors, e.g., nested Chinese restaurant processes (Blei et al., 2004)
- **Learning hyperparameters from data**
- **Identifiability and interpretation/evaluation of results**
- **Discriminative tasks** (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008; Mairal et al., 2009d)
- **Optimization and local minima**

Structured sparsity - Audio processing
Source separation (Lefèvre et al., 2011)



Structured sparsity - Audio processing
Musical instrument separation (Lefèvre et al., 2011)

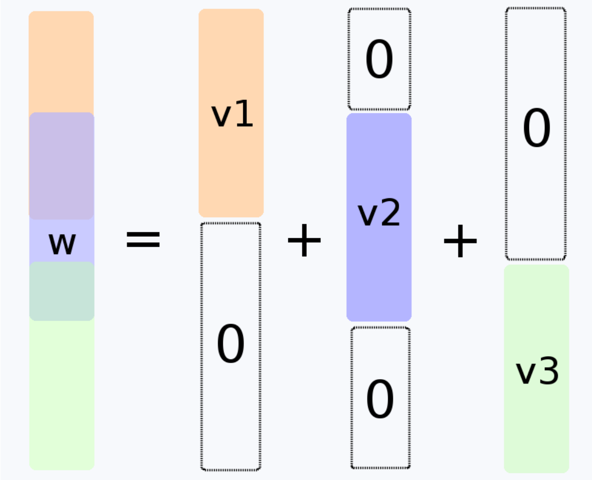
- Unsupervised source separation with group-sparsity prior
  - Top: mixture
  - Left: source tracks (guitar, voice). Right: separated tracks.



Alternative approach: latent group Lasso (Jacob, Obozinski, and Vert, 2009)

- **Overlapping I:** $\Omega(w) = \sum_{G \in \mathbf{G}} \|w_G\|_2$
  - Sparsity patterns invariant by intersection
- **Overlapping II:** $\Omega(w) = \inf_{w=\sum_{G \in \mathbf{G}} v_G, \ \mathrm{Supp}(v_G) \subseteq G} \sum_{G \in \mathbf{G}} \|v_G\|_2$

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{G \in \mathbf{G}} \|v_G\|_2 \\ w = \sum_{G \in \mathbf{G}} v_G \\ \mathrm{Supp}(v_G) \subseteq G \end{cases}$$



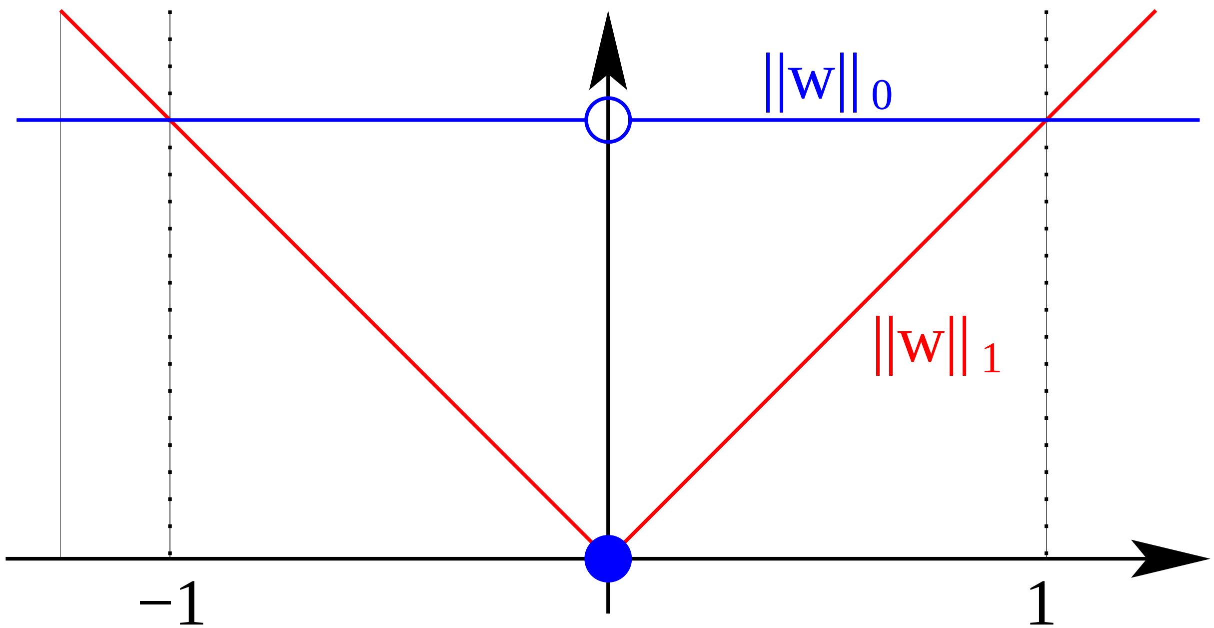
  - Sparsity patterns invariant by union

Outline

- **Tutorial: Sparse methods for machine learning**
  - Algorithms: Convex optimization
  - Theory: high-dimensional inference
  - Learning on matrices
- **Classical approaches to structured sparsity**
  - Linear combinations of ℓ_q-norms
  - Applications
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis

ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\mathrm{Supp}(w) = \{j \in V, \ w_j \neq 0\}$
- **Cardinality of support**: $\|w\|_0 = \mathrm{Card}(\mathrm{Supp}(w))$
- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)



- ℓ_1-norm = convex envelope of ℓ_0-quasi-norm on the ℓ_∞-ball $[-1, 1]^p$

Convex envelopes of general functions of the support (Bach, 2010)

- Let $F : 2^V \to \mathbb{R}$ be a **set-function**
  - Assume F is **non-decreasing** (i.e., $A \subset B \Rightarrow F(A) \leqslant F(B)$)
  - Explicit prior knowledge on supports (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)
- Define $\Theta(w) = F(\text{Supp}(w))$: How to get its convex envelope?
  1. Possible if F is also **submodular**
  2. Allows **unified** theory and algorithm
  3. Provides **new** regularizers

Submodular functions (Fujishige, 2005; Bach, 2011)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$
\begin{aligned}
& \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\
\Leftrightarrow \ & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}
\end{aligned}
$$

Submodular functions (Fujishige, 2005; Bach, 2011)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \ & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

Submodular functions (Fujishige, 2005; Bach, 2011)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \quad & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  – Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

- **Intuition 2**: behave like convex functions
  – Polynomial-time minimization, conjugacy theory

Submodular functions (Fujishige, 2005; Bach, 2011)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$

$$\Leftrightarrow \quad \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  – Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

- **Intuition 2**: behave like convex functions
  – Polynomial-time minimization, conjugacy theory

- Used in several areas of signal processing and machine learning
  – Total variation/graph cuts (Chambolle, 2005; Boykov et al., 2001)
  – Optimal design (Krause and Guestrin, 2005)

Submodular functions - Examples

- Concave functions of the cardinality: $g(|A|)$
- Cuts
- Entropies
  - $H((X_k)_{k \in A})$ from p random variables $X_1, \dots, X_p$
- Network flows
  - Efficient representation for set covers
- Rank functions of matroids

Submodular functions - Lovász extension

- Subsets may be identified with elements of $\{0,1\}^p$
- Given any set-function F and w such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, define:

$$f(w) = \sum_{k=1}^{p} w_{j_k} [F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})]$$

  - If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  - f is piecewise affine and positively homogeneous

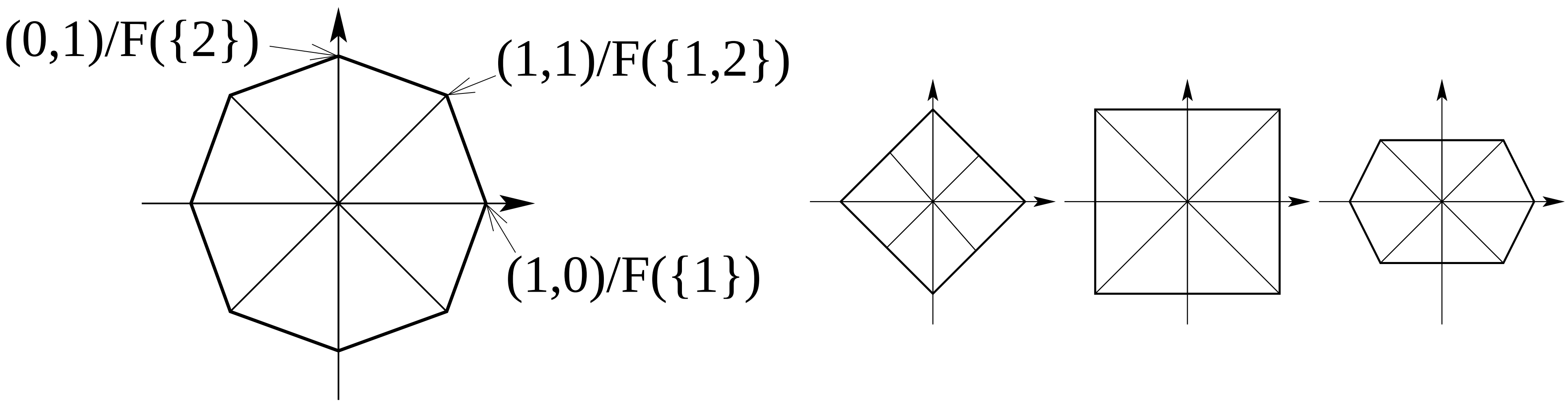
- F is submodular if and only if f is convex (Lovász, 1982)
  - Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V

Submodular functions and structured sparsity

- Let $F : 2^V \rightarrow \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

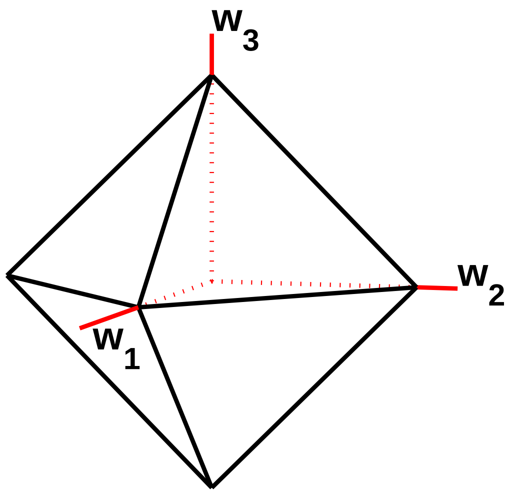
Submodular functions and structured sparsity

- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F
- **Sparsity-inducing properties**: Ω is a polyhedral norm

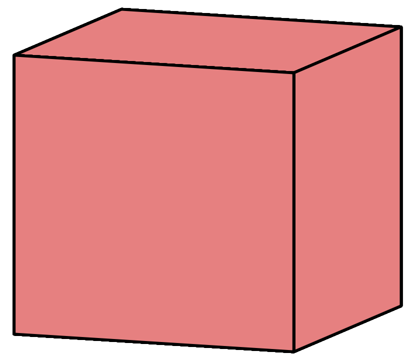


  - A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  - With probability one, stable sets are the only allowed active sets

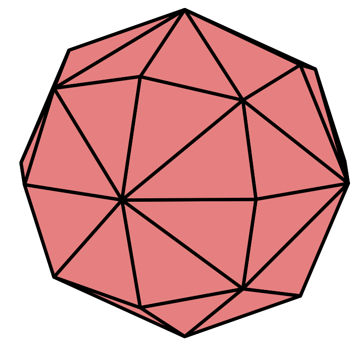
Polyhedral unit balls



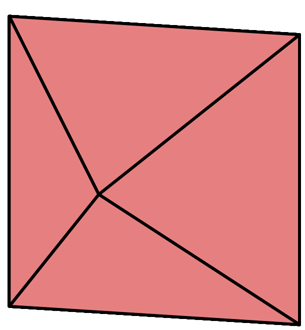
$F(A) = |A|$
$\Omega(w) = \|w\|_1$



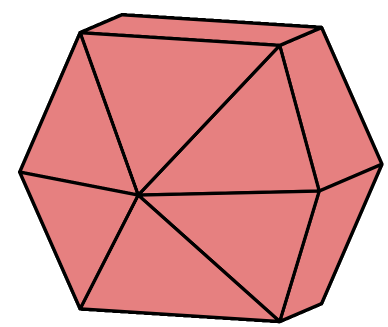
$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



$F(A) = |A|^{1/2}$
all possible extreme points



$F(A) = 1_{\{A \cap \{1\} \neq \varnothing\}} + 1_{\{A \cap \{2,3\} \neq \varnothing\}}$
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$



$F(A) = 1_{\{A \cap \{1,2,3\} \neq \varnothing\}}$
$+ 1_{\{A \cap \{2,3\} \neq \varnothing\}} + 1_{\{A \cap \{3\} \neq \varnothing\}}$
$\Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3|$

Submodular functions and structured sparsity

- **Unified theory and algorithms**
  - Generic computation of proximal operator
  - Unified oracle inequalities

- **Extensions**
  - Shaping level sets through symmetric submodular function (Bach, 2011)
  - ℓ_q-relaxations of combinatorial penalties (Obozinski and Bach, 2011)

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty$$

  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

$$\big(\mathrm{Supp}(w)\big)^c = \bigcup_{G \in \mathbf{H}'} G \ \text{ for some } \mathbf{H}' \subseteq \mathbf{H}$$

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \;\Rightarrow\quad F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\; G \cap A \neq \varnothing\}\big)$$

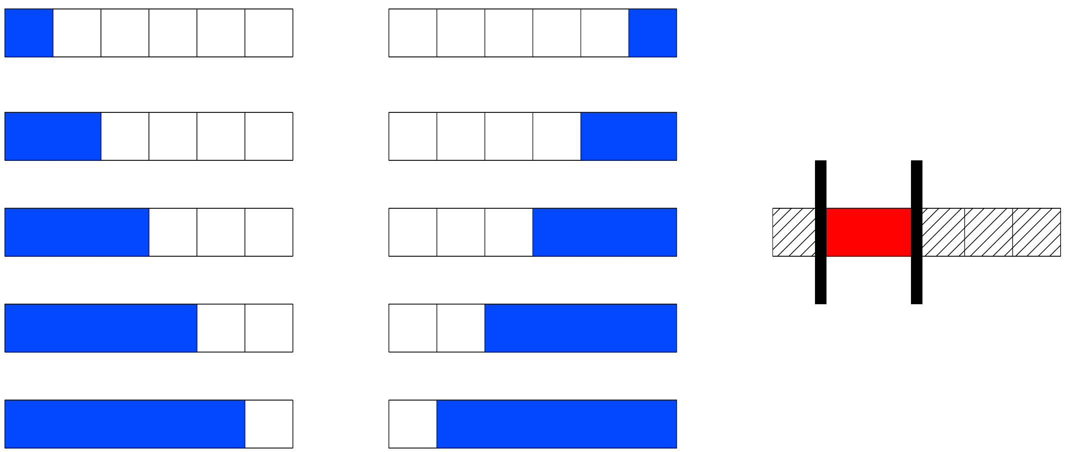
  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

$$\big(\mathrm{Supp}(w)\big)^c = \bigcup_{G \in \mathbf{H}} G \quad \text{for some } \mathbf{H}' \subseteq \mathbf{H}$$

  - Justification not only limited to allowed sparsity patterns

Selection of contiguous patterns in a sequence

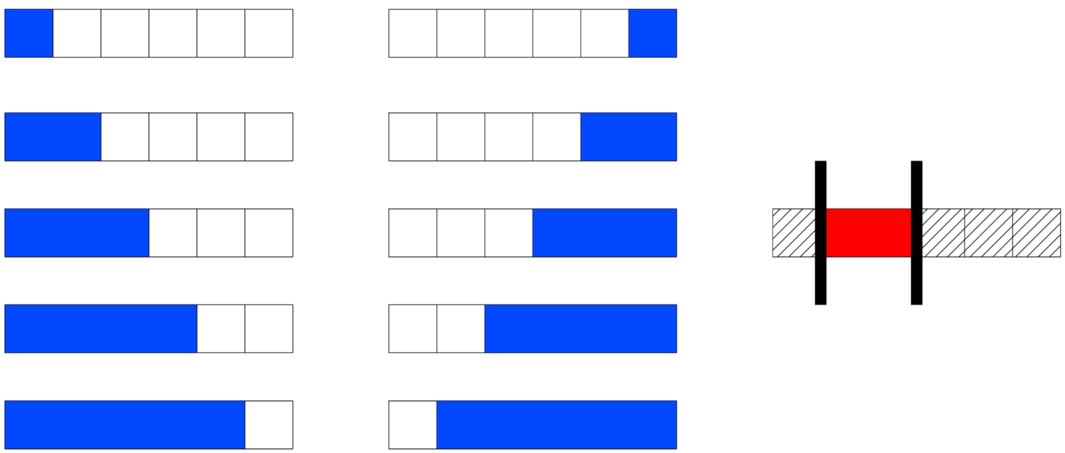
- Selection of contiguous patterns in a sequence



- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

Selection of contiguous patterns in a sequence

- Selection of contiguous patterns in a sequence

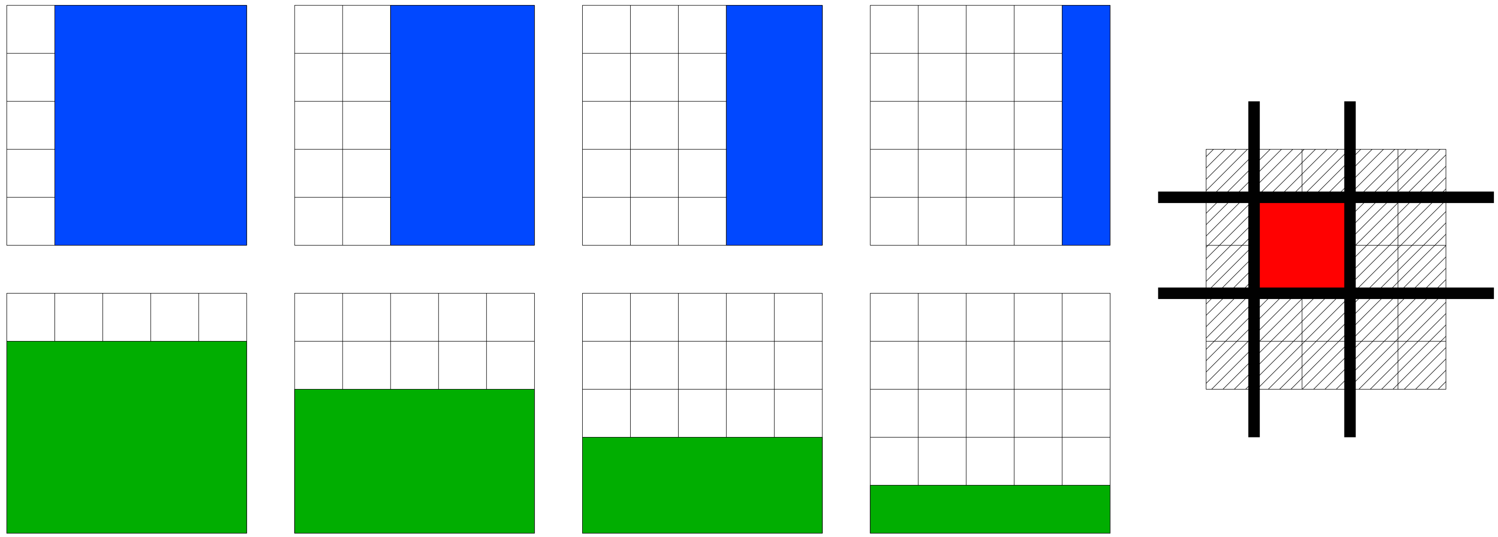


- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

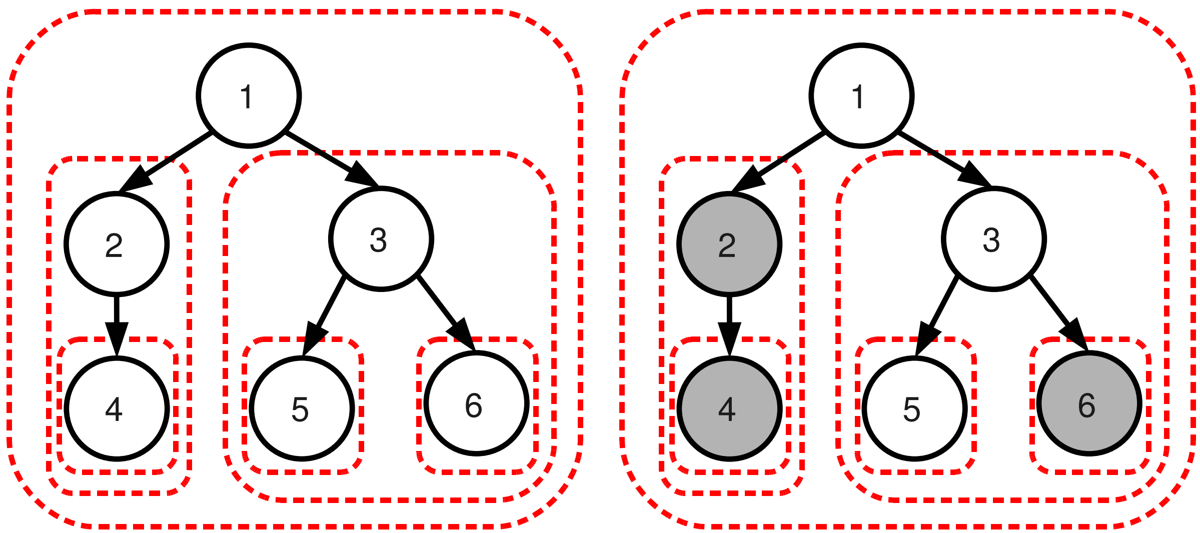
- $\sum_{G\in\mathbf{H}} \|w_G\|_\infty \;\Rightarrow\; F(A) = p - 1 + \mathrm{Range}(A)$ if $A \neq \varnothing$
  - Jump from 0 to $p-1$: tends to include all variables simultaneously
  - Add $\nu|A|$ to smooth the kink: all sparsity patterns are possible
  - **Contiguous patterns are favored (and not forced)**

Extensions of norms with overlapping groups

- Selection of **rectangles** (at any position) in a 2-D grids



- **Hierarchies**



Submodular functions and structured sparsity
Examples

- **From** $\Omega(w)$ **to** $F(A)$: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\ G \cap A \neq \varnothing\}\big)$$

  - Justification not only limited to allowed sparsity patterns

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\ G \cap A \neq \varnothing\}\big)$$

  - Justification not only limited to allowed sparsity patterns

- **From $F(A)$ to $\Omega(w)$:** provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \ \Rightarrow\ \Omega$ is a combination of **order statistics**
  - **Non-factorial priors** for supervised learning: Ω depends on the eigenvalues of $X_A^\top X_A$ and not simply on the cardinality of A

Non-factorial priors for supervised learning

- **Joint variable selection and regularization**. Given support $A \subset V$,

$$\min_{w_A \in \mathbb{R}^A} \frac{1}{2n}\|y - X_A w_A\|_2^2 + \frac{\lambda}{2}\|w_A\|_2^2$$

- Minimizing with respect to A will always lead to $A = V$

- **Information/model selection criterion** $F(A)$

$$\min_{A \subset V} \min_{w_A \in \mathbb{R}^A} \frac{1}{2n}\|y - X_A w_A\|_2^2 + \frac{\lambda}{2}\|w_A\|_2^2 + F(A)$$

$$\Leftrightarrow \quad \min_{w \in \mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 + F(\mathrm{Supp}(w))$$

Non-factorial priors for supervised learning

- Selection of subset A from design $X \in \mathbb{R}^{n\times p}$ with ℓ_2-penalization
- **Frequentist analysis** (Mallow's C_L): $\operatorname{tr} X_A^\top X_A (X_A^\top X_A + \lambda I)^{-1}$
  - Not submodular
- **Bayesian analysis** (marginal likelihood): $\log\det(X_A^\top X_A + \lambda I)$
  - Submodular (also true for $\operatorname{tr}(X_A^\top X_A)^{1/2}$)

p	n	k	submod.	ℓ_2 vs. submod.	ℓ_1 vs. submod.	greedy vs. submod.
120	120	80	40.8 $\pm$ 0.8	-2.6 $\pm$ 0.5	**0.6 $\pm$ 0.0**	**21.8 $\pm$ 0.9**
120	120	40	35.9 $\pm$ 0.8	**2.4 $\pm$ 0.4**	**0.3 $\pm$ 0.0**	**15.8 $\pm$ 1.0**
120	120	20	29.0 $\pm$ 1.0	**9.4 $\pm$ 0.5**	-0.1 $\pm$ 0.0	**6.7 $\pm$ 0.9**
120	120	10	20.4 $\pm$ 1.0	**17.5 $\pm$ 0.5**	-0.2 $\pm$ 0.0	-2.8 $\pm$ 0.8
120	20	20	49.4 $\pm$ 2.0	0.4 $\pm$ 0.5	**2.2 $\pm$ 0.8**	**23.5 $\pm$ 2.1**
120	20	10	49.2 $\pm$ 2.0	0.0 $\pm$ 0.6	1.0 $\pm$ 0.8	**20.3 $\pm$ 2.6**
120	20	6	43.5 $\pm$ 2.0	**3.5 $\pm$ 0.8**	**0.9 $\pm$ 0.6**	**24.4 $\pm$ 3.0**
120	20	4	41.0 $\pm$ 2.1	**4.8 $\pm$ 0.7**	-1.3 $\pm$ 0.5	**25.1 $\pm$ 3.5**

Unified optimization algorithms

- **Polyhedral norm** with $O(3^p)$ faces and extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow

Unified optimization algorithms

- **Polyhedral norm** with $O(3^p)$ faces and extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow
- **Proximal methods** (e.g., Beck and Teboulle, 2009)
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when (P) : $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda\Omega(w)$ is "easy"
- **Proposition**: (P) is equivalent to submodular function minimization

Proximal methods for Lovász extensions

- **Proposition** (Chambolle and Darbon, 2009): let w^* be the solution of $\min_{w\in\mathbb{R}^p} \frac{1}{2}\|w-v\|_2^2 + \lambda f(w)$. Then the solutions of

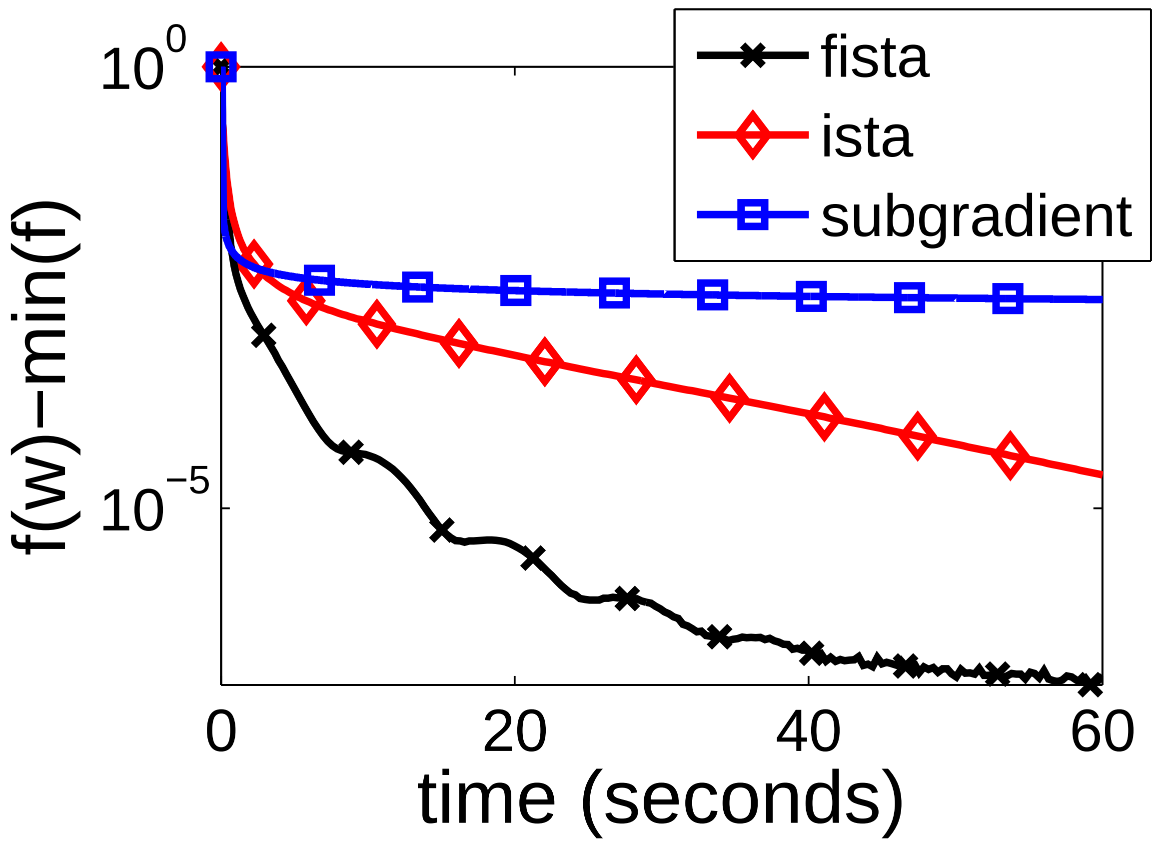
$$\min_{A\subset V} \lambda F(A) + \sum_{j\in A}(\alpha - v_j)$$

  are the sets A^α such that $\{w^* > \alpha\} \subset A^\alpha \subset \{w^* \geqslant \alpha\}$

- **Parametric submodular function optimization**
  - General decomposition strategy for $f(|w|)$ and $f(w)$ (Groenevelt, 1991)
  - Efficient only when submodular minimization is efficient
  - Otherwise, minimum-norm-point algorithm (a.k.a. Frank Wolfe) is preferable

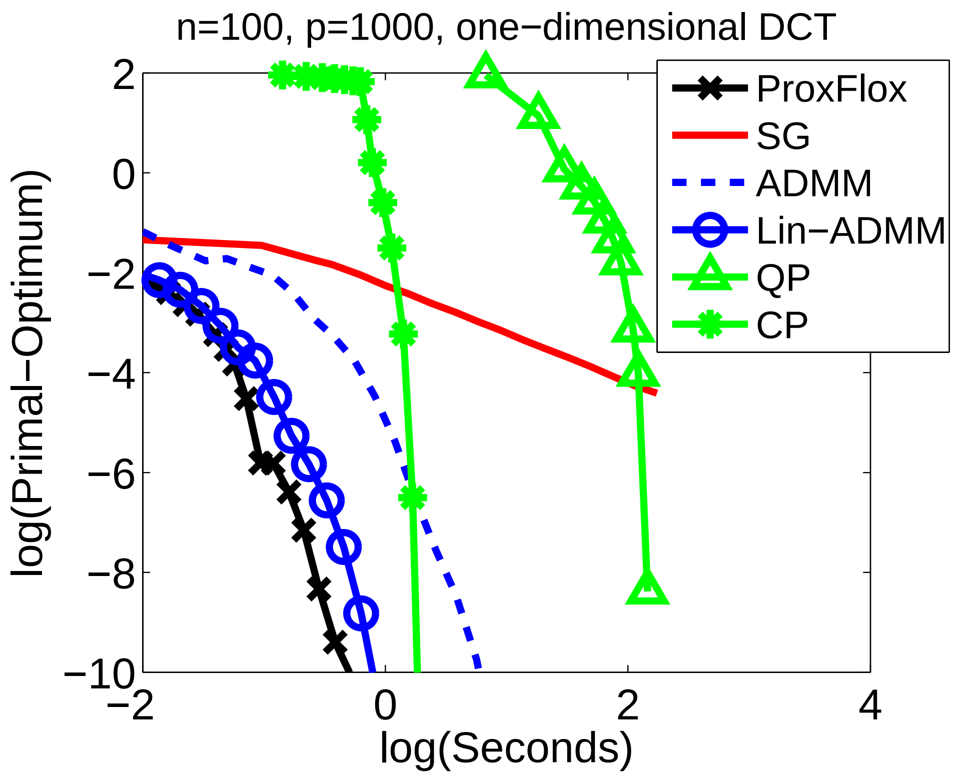
Comparison of optimization algorithms

- Synthetic example with $p = 1000$ and $F(A) = |A|^{1/2}$
- ISTA: proximal method
- FISTA: accelerated variant (Beck and Teboulle, 2009)



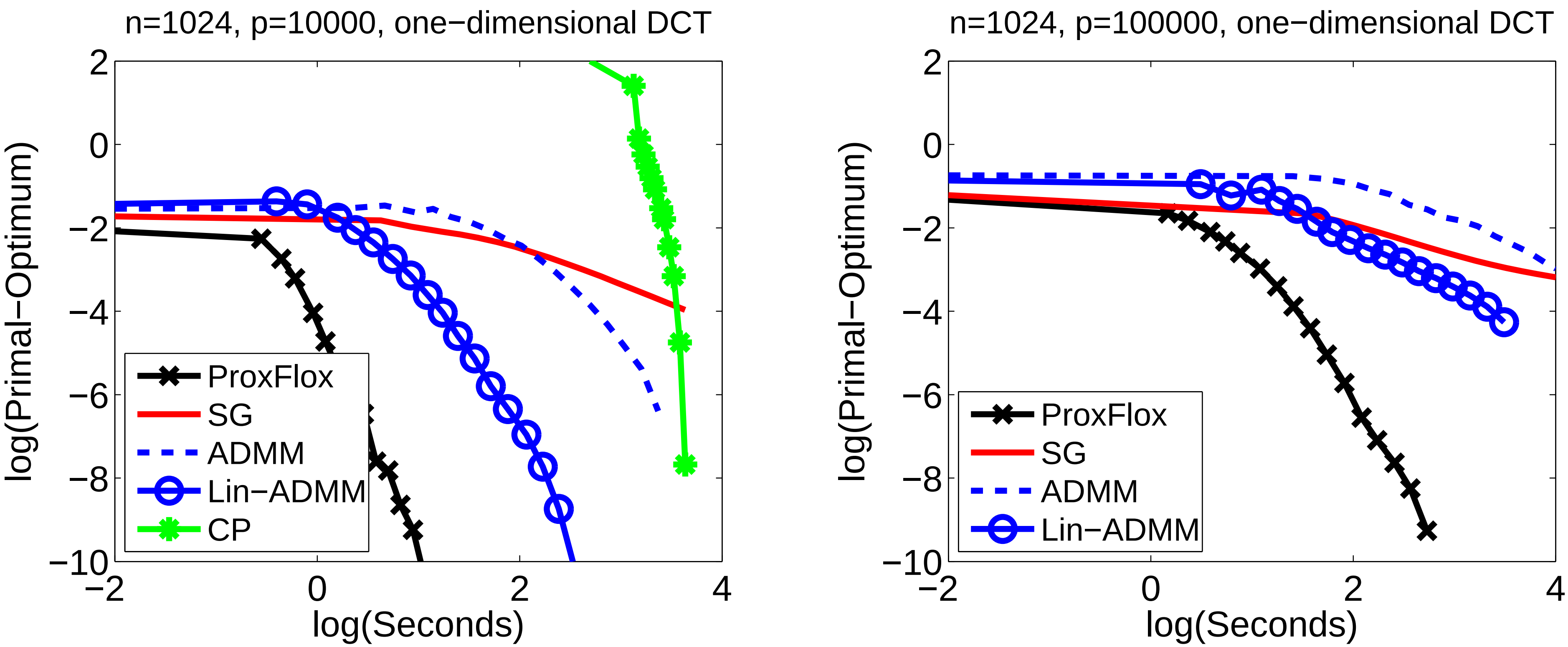
Comparison of optimization algorithms (Mairal, Jenatton, Obozinski, and Bach, 2010b) Small scale

- Specific norms which can be implemented through network flows



Comparison of optimization algorithms (Mairal, Jenatton, Obozinski, and Bach, 2010b)
Large scale

- Specific norms which can be implemented through network flows



Unified theoretical analysis

- **Decomposability**
  - Key to theoretical analysis (Negahban et al., 2009)
  - **Property**: $\forall w \in \mathbb{R}^p$, and $\forall J \subset V$, if $\min_{j \in J} |w_j| \geqslant \max_{j \in J^c} |w_j|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$

- **Support recovery**
  - Extension of known sufficient condition (Zhao and Yu, 2006; Negahban and Wainwright, 2008)

- **High-dimensional inference**
  - Extension of known sufficient condition (Bickel et al., 2009)
  - Matches with analysis of Negahban et al. (2009) for common cases

Support recovery - $\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$

- **Notation**
  - $\rho(J) = \min_{B\subset J^c} \frac{F(B\cup J)-F(J)}{F(B)} \in (0,1]$ (for J stable)
  - $c(J) = \sup_{w\in\mathbb{R}^p} \Omega_J(w_J)/\|w_J\|_2 \leqslant |J|^{1/2} \max_{k\in V} F(\{k\})$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$
  - J = smallest stable set containing the support of w^*
  - Assume $\nu = \min_{j, w_j^*\neq 0} |w_j^*| > 0$
  - Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p\times p}$. Assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$
  - Assume that for $\eta > 0$, $\boxed{(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1}Q_{Jj}))_{j\in J^c}] \leqslant 1 - \eta}$
  - If $\lambda \leqslant \frac{\kappa\nu}{2c(J)}$, $\hat{w}$ has support equal to J, with probability larger than

$$1 - 3P\Big(\Omega^*(z) > \frac{\lambda\eta\rho(J)\sqrt{n}}{2\sigma}\Big)$$

  - z is a multivariate normal with covariance matrix Q

Consistency - $\min_{w\in\mathbb{R}^p} \ \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$
  - J = smallest stable set containing the support of w^*
  - Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p\times p}$.
  - Assume that $\forall\Delta$ s.t. $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, $\Delta^\top Q\Delta \geqslant \kappa\|\Delta_J\|_2^2$
  - Then $\boxed{\Omega(\hat{w} - w^*) \leqslant \dfrac{24c(J)^2\lambda}{\kappa\rho(J)^2}}$ and $\boxed{\dfrac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \dfrac{36c(J)^2\lambda^2}{\kappa\rho(J)^2}}$
    with probability larger than $1 - P\big(\Omega^*(z) > \frac{\lambda\rho(J)\sqrt{n}}{2\sigma}\big)$
  - z is a multivariate normal with covariance matrix Q

- **Concentration inequality** (z normal with covariance matrix Q):
  - $\mathcal{T}$ set of stable inseparable sets
  - Then $P(\Omega^*(z) > t) \leqslant \sum_{A\in\mathcal{T}} 2^{|A|} \exp\big(- \frac{t^2F(A)^2/2}{1^\top Q_{AA}1}\big)$

Symmetric submodular functions (Bach, 2011)

- Let $F : 2^V \to \mathbb{R}$ be a **symmetric submodular set-function**

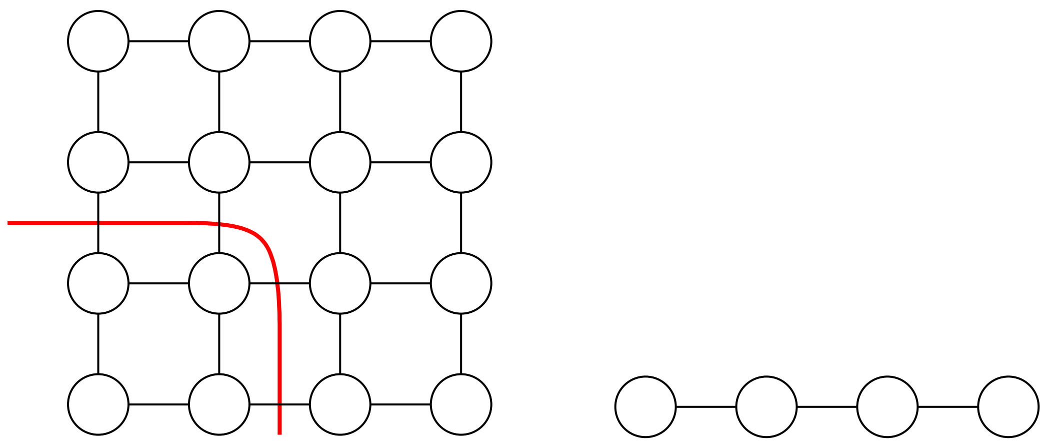
- **Proposition**: The Lovász extension $f(w)$ is the convex envelope of the function $w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$ on the set $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p,\ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.

- **Shaping all level sets**

Symmetric submodular functions - Examples

- **From $\Omega(w)$ to $F(A)$**: provides new insights into existing norms
  - **Cuts - total variation**

$$F(A) = \sum_{k \in A, j \in V \backslash A} d(k,j) \quad \Rightarrow \quad f(w) = \sum_{k,j \in V} d(k,j)(w_k - w_j)_+$$

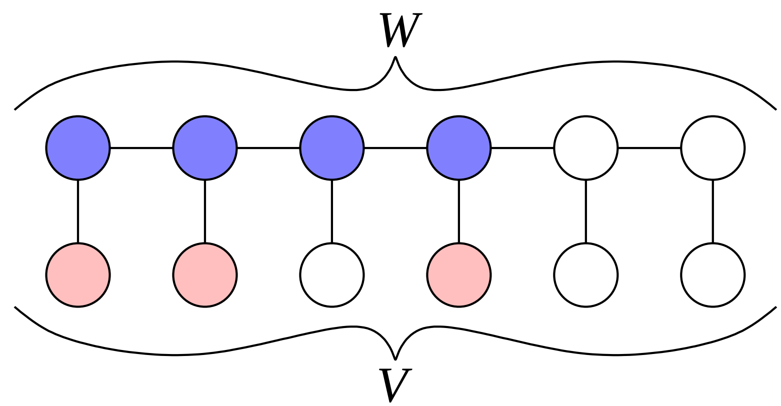


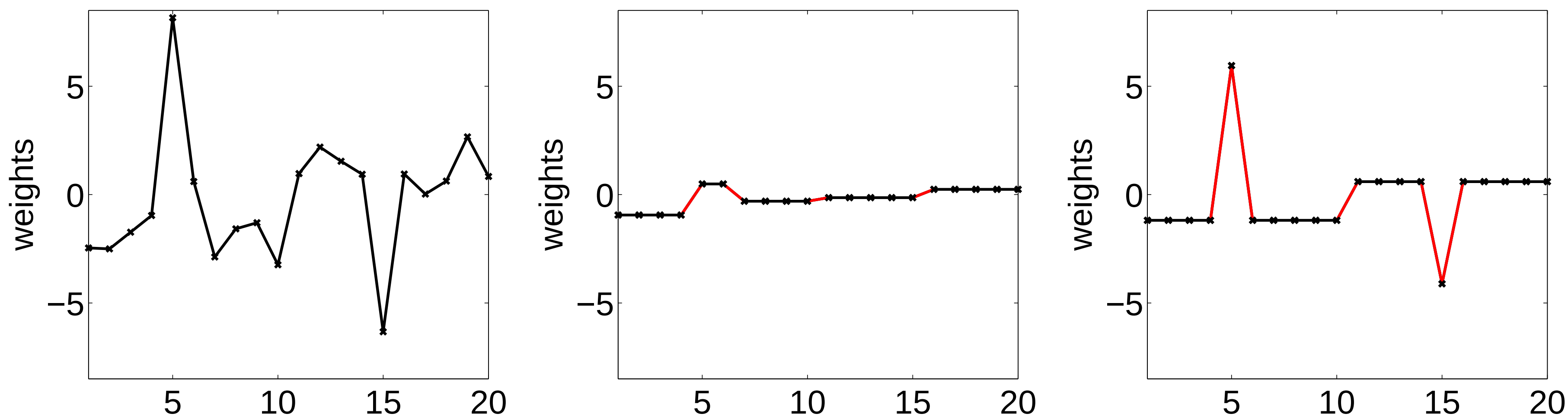
  - NB: graph may be directed
  - Application to change-point detection (Tibshirani et al., 2005; Harchaoui and Lévy-Leduc, 2008)

Symmetric submodular functions - Examples

- **From $F(A)$ to $\Omega(w)$**: provides new sparsity-inducing norms
  - **Regular functions** (Boykov et al., 2001; Chambolle and Darbon, 2009)

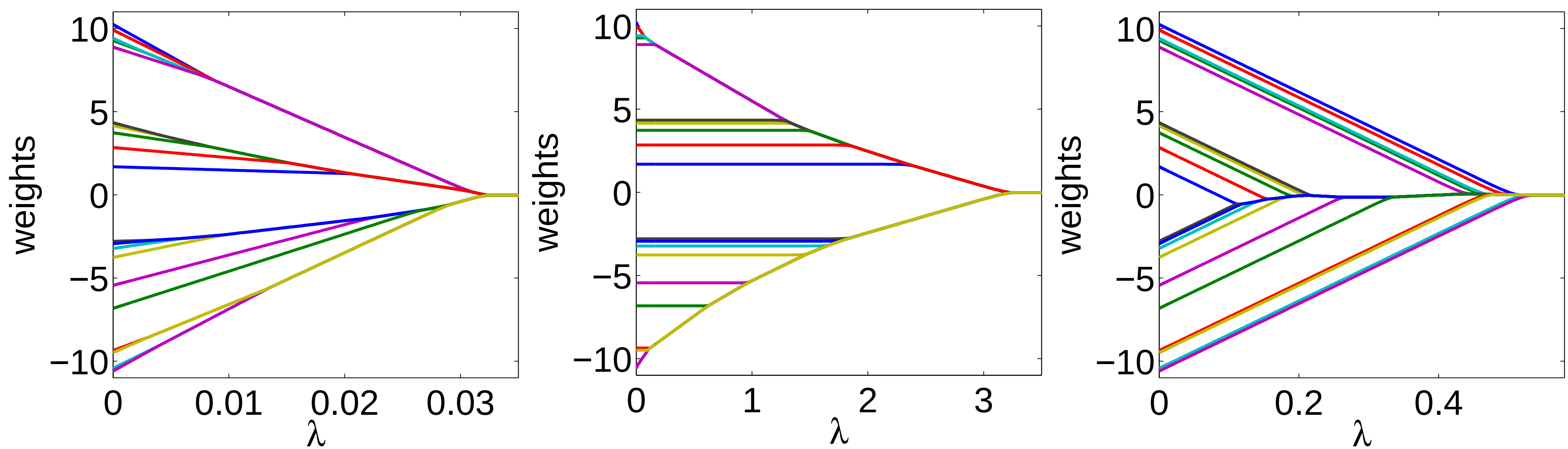
$$F(A) = \min_{B \subset W} \sum_{k \in B,\ j \in W \setminus B} d(k,j) + \lambda |A \Delta B|$$





Symmetric submodular functions - Examples

- **From $F(A)$ to $\Omega(w)$:** provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \Rightarrow$ priors on the size and numbers of clusters

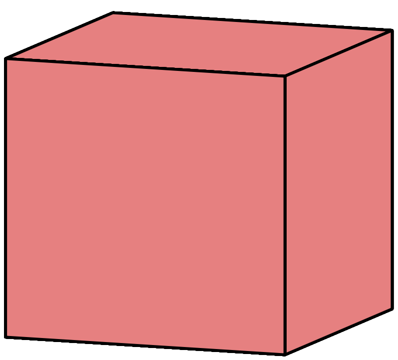


$|A|(p-|A|)$ $\quad$ $1_{|A|\in(0,p)}$ $\quad$ $\max\{|A|, p-|A|\}$

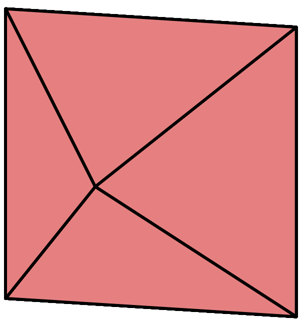
  - Convex formulations for clustering (Hocking, Joulin, Bach, and Vert, 2011)

ℓ_q-relaxation of combinatorial penalties (Obozinski and Bach, 2011)

- **Main result** of Bach (2010):
  - $f(|w|)$ is the convex envelope of $F(\mathrm{Supp}(w))$ on $[-1,1]^p$
- **Problems**:
  - Limited to submodular functions
  - Limited to ℓ_∞-relaxation: undesired artefacts



$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



$F(A) = 1_{\{A \cap \{1\} \neq \varnothing\}} + 1_{\{A \cap \{2,3\} \neq \varnothing\}}$
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$

From ℓ_∞ to ℓ_2

- Variational formulations for subquadratic norms (Bach et al., 2011)

$$\Omega(w) = \min_{\eta \in \mathbb{R}_+^p} \frac{1}{2} \sum_{j=1}^p \frac{w_j^2}{\eta_j} + \frac{1}{2} g(\eta) = \min_{\eta \in H} \sqrt{\sum_{j=1}^p \frac{w_j^2}{\eta_j}}$$

  where g is a convex homogeneous and $H = \{\eta, g(\eta) \leqslant 1\}$

  – Often used for computational reasons (Lasso, group Lasso)
  – May also be used to define a norm (Micchelli et al., 2011)

From ℓ_∞ to ℓ_2

- Variational formulations for subquadratic norms (Bach et al., 2011)

$$\Omega(w) = \min_{\eta \in \mathbb{R}^p_+} \frac{1}{2} \sum_{j=1}^p \frac{w_j^2}{\eta_j} + \frac{1}{2} g(\eta) = \min_{\eta \in H} \sqrt{\sum_{j=1}^p \frac{w_j^2}{\eta_j}}$$

  where g is a convex homogeneous and $H = \{\eta, g(\eta) \leqslant 1\}$

  – Often used for computational reasons (Lasso, group Lasso)
  – May also be used to define a norm (Micchelli et al., 2011)

- If F is a nondecreasing submodular function with Lovász extension f

  – Define $\Omega_2^F(w) = \min_{\eta \in \mathbb{R}^p_+} \frac{1}{2} \sum_{j=1}^p \frac{w_j^2}{\eta_j} + \frac{1}{2} f(\eta)$
  – Is it the convex relaxation of some natural function?

ℓ_q-relaxation of submodular penalties (Obozinski and Bach, 2011)

- F a nondecreasing submodular function with Lovász extension f
- Define $\Omega_q^F(w) = \min_{\eta \in \mathbb{R}_+^p} \dfrac{1}{q} \sum_{i \in V} \dfrac{|w_i|^q}{\eta_i^{q-1}} + \dfrac{1}{r} f(\eta)$ with $\dfrac{1}{q} + \dfrac{1}{r} = 1$
- **Proposition 1**: Ω_q^F is the convex envelope of $w \mapsto F(\mathrm{Supp}(w)) \|w\|_q$
- **Proposition 2**: Ω_q^F is the homogeneous convex envelope of $w \mapsto \frac{1}{r} F(\mathrm{Supp}(w)) + \frac{1}{q} \|w\|_q^q$
- **Jointly penalizing and regularizing**
  - Special cases $q = 1$, $q = 2$ and $q = \infty$
- Removes artefacts of ℓ_∞-formulation

How tight is the relaxation?

What information of F is kept after the relaxation?

- When F is **submodular** and $q = \infty$
  - the Lovász extension $f = \Omega_\infty^F$ is said to "extend" F because $\Omega_\infty^F(1_A) = f(1_A) = F(A)$

- **In general** we can still consider the function : $G(A) \triangleq \Omega_\infty^F(1_A)$
  - Do we have $G(A) = F(A)$?
  - How is G related to F?
  - What is the norm Ω_∞^G which is associated with G?

Lower combinatorial envelope

- Given a function $F : 2^V \to \mathbb{R}$, define its lower combinatorial envelope as the function G given by

$$G(A) = \max_{s \in P(F)} s(A)$$

with $P(F) = \{s \in \mathbb{R}^p,\ \forall A \subset V,\ s(A) \leq F(A)\}$.

- **Property 1** : G is the largest function such that $G \leqslant F$ and

$$G(A) = \Omega_\infty^G(1_A)$$

- **Property 2** : G is its own combinatorial envelope

- **A new class of set-functions**

Conclusion

- **Structured sparsity for machine learning and statistics**
  - Many applications (image, audio, text, etc.)
  - May be achieved through structured sparsity-inducing norms
  - Link with submodular functions: unified analysis and algorithms

    Submodular functions to encode discrete structures

Conclusion

- **Structured sparsity for machine learning and statistics**
  - Many applications (image, audio, text, etc.)
  - May be achieved through structured sparsity-inducing norms
  - Link with submodular functions: unified analysis and algorithms

    Submodular functions to encode discrete structures

- **On-going work on structured sparsity**
  - Norm design beyond submodular functions
  - Instance of general framework of Chandrasekaran et al. (2010)
  - Links with greedy (i.e., non convex) methods (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)
  - Achieving $\log p = O(n)$ algorithmically (Bach, 2008c)

References

Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine Learning (ICML)*, 2007.

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. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.

F. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. 2011. URL `http://hal.inria.fr/hal-00645271/en`.

F. Bach. Shaping level sets with submodular functions. In *Adv. NIPS*, 2011.

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)*, 2004.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv, 2008.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report 00613125, HAL, 2011.

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.

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

L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Speech and Audio Processing*, 14(1):191, 2006.

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.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, January 2003.

D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

D.M. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

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.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

A. Buades, B. Coll, and J.-M. Morel. Non-local image and movie denoising. *International Journal of Computer vision*, 76(2):123–139, 2008.

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.

E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.

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

V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems*, 2008.

A. Chambolle. Total variation minimization and a class of binary MRF models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer, 2005.

A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Arxiv preprint arXiv:1012.0621*, 2010.

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

Florent Couzinie-Devy, Julien Mairal, Francis Bach, and Jean Ponce. Dictionary Learning for Deblurring and Digital Zoom. Technical report, September 2011. URL `http://hal.inria.fr/inria-00627402`.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

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.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

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.

M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734–4739, 2001.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis. *Neural Computation*, 21(3), 2009.

J. Friedman, T. Hastie, H. Höfling, 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).

S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.

H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.

Z. Harchaoui and C. Lévy-Leduc. Catching change-points with Lasso. *Adv. NIPS*, 20, 2008.

Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale classification with trace-norm regularization. In *Proc. CVPR*, 2012.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proc. ICML*, 2011.

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.

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

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

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

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

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

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multi-scale mining of fmri data with hierarchical structured sparsity. Technical report, Preprint arXiv:1105.0363,

2011. In submission to SIAM Journal on Imaging Sciences.

K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of CVPR*, 2009.

S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.

S. Lacoste-Julien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems (NIPS) 21*, 2008.

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

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

K. Lounici, A.B. Tsybakov, M. Pontil, and S.A. van de Geer. Taking advantage of sparsity in multi-task learning. In *Conference on Computational Learning Theory (COLT)*, 2009.

L. Lovász. Submodular functions and convexity. *Mathematical programming: the state of the art, Bonn*, pages 235–257, 1982.

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.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. Technical report, arXiv:0908.0050, 2009a.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, 2009b.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009c.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 21, 2009d.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *International Conference on Computer Vision (ICCV)*, 2009e.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *NIPS*, 2010b.

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.

C.A. Micchelli, J.M. Morales, and M. Pontil. Regularizers for structured sparsity. *Arxiv preprint arXiv:1010.0556*, 2011.

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

S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1-ℓ_∞-regularization. In *Adv. NIPS*, 2008.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2009.

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 and F. Bach. Convex relaxation of combinatorial penalties. Technical report, HAL, 2011.

G. Obozinski, M.J. Wainwright, and M.I. Jordan. High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

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.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

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

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 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.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

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

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *J. Roy. Stat. Soc. B*, 67(1):91–108, 2005.

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

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse structured dictionary learning for brain resting-state activity modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

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.

D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 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.