Structured sparsity through convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

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

Cambridge University - May 2012

Outline

- **Introduction: Sparse methods for machine learning**
  - **Short tutorial**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **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:

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

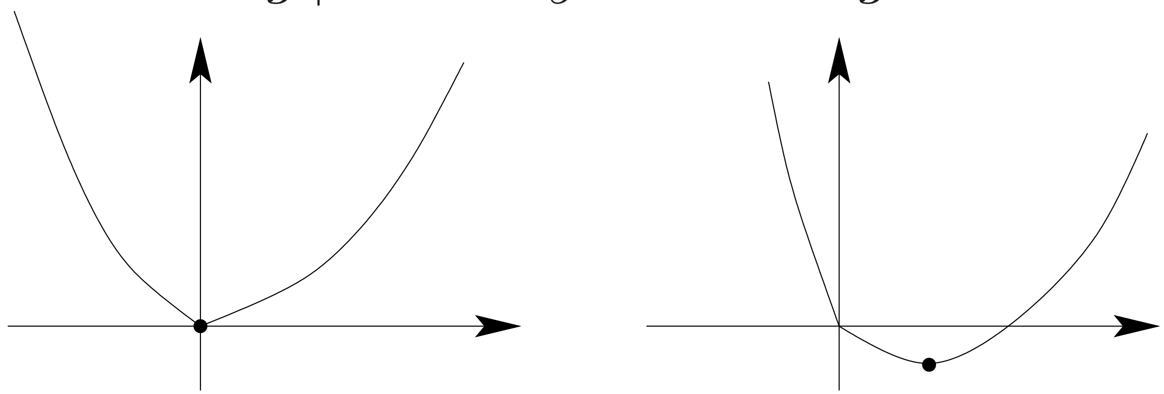
- Norm Ω to promote sparsity
  - square loss + ℓ_1-norm $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$

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 $0-$: $g_- = -\lambda - y$

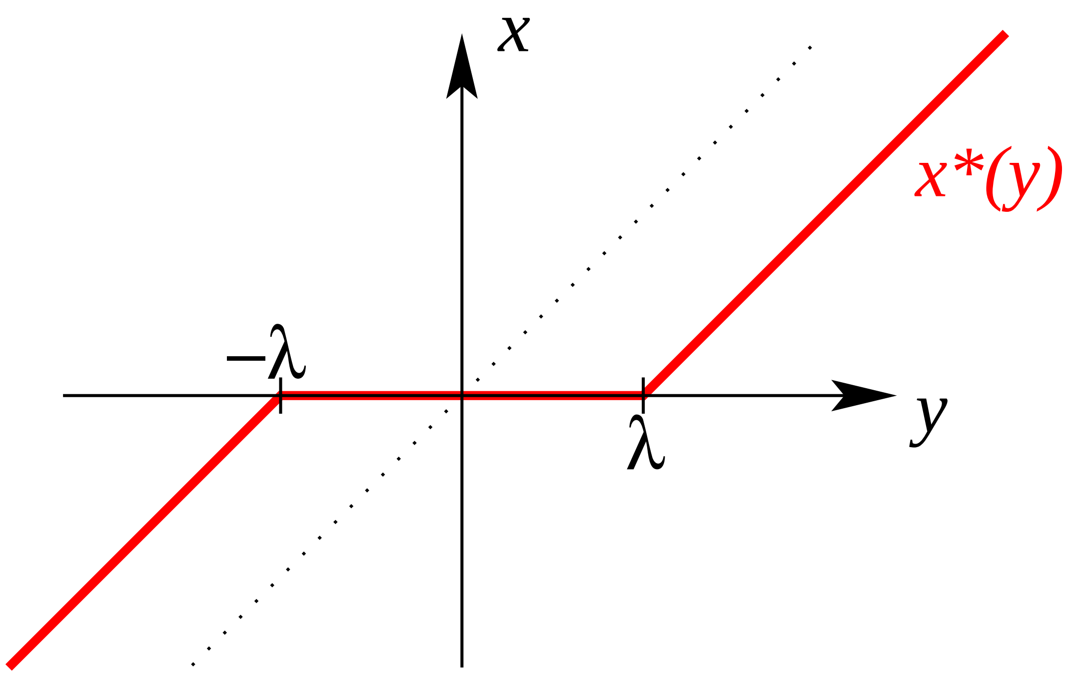


  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \leqslant 0$ (i.e., $y \leqslant -\lambda$)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

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

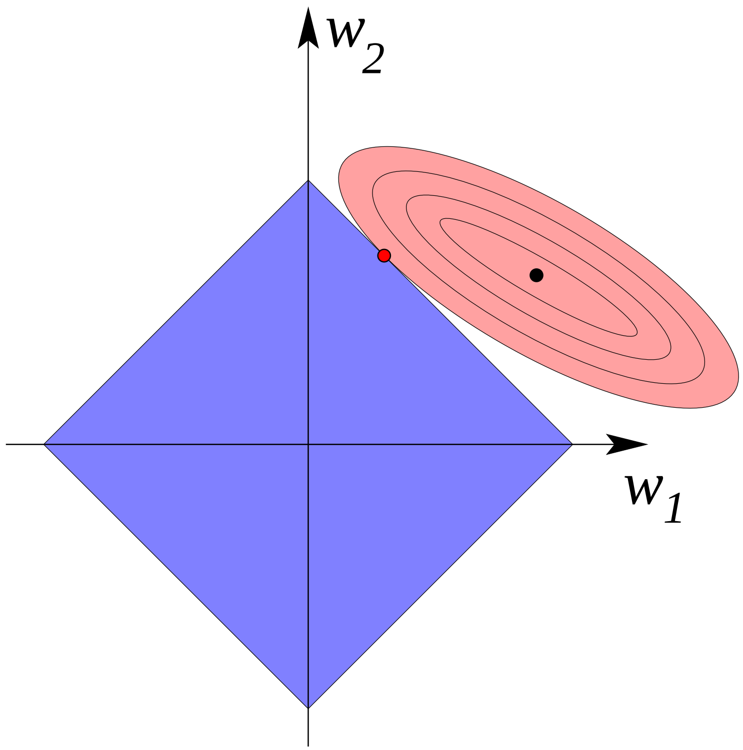
Why ℓ_1-norms lead to sparsity?

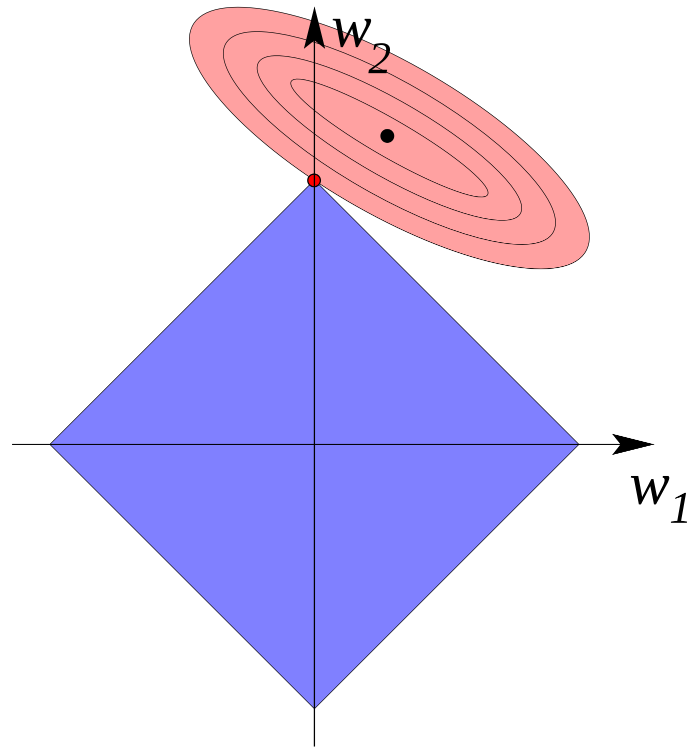
- **Example 1**: quadratic problem in 1D, i.e. $\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|$
- Piecewise quadratic function with a kink at zero
- Solution $x^* = \text{sign}(y)(|y| - \lambda)_+$ = soft thresholding



Why ℓ_1-norms lead to sparsity?

- **Example 2**: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding
- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining



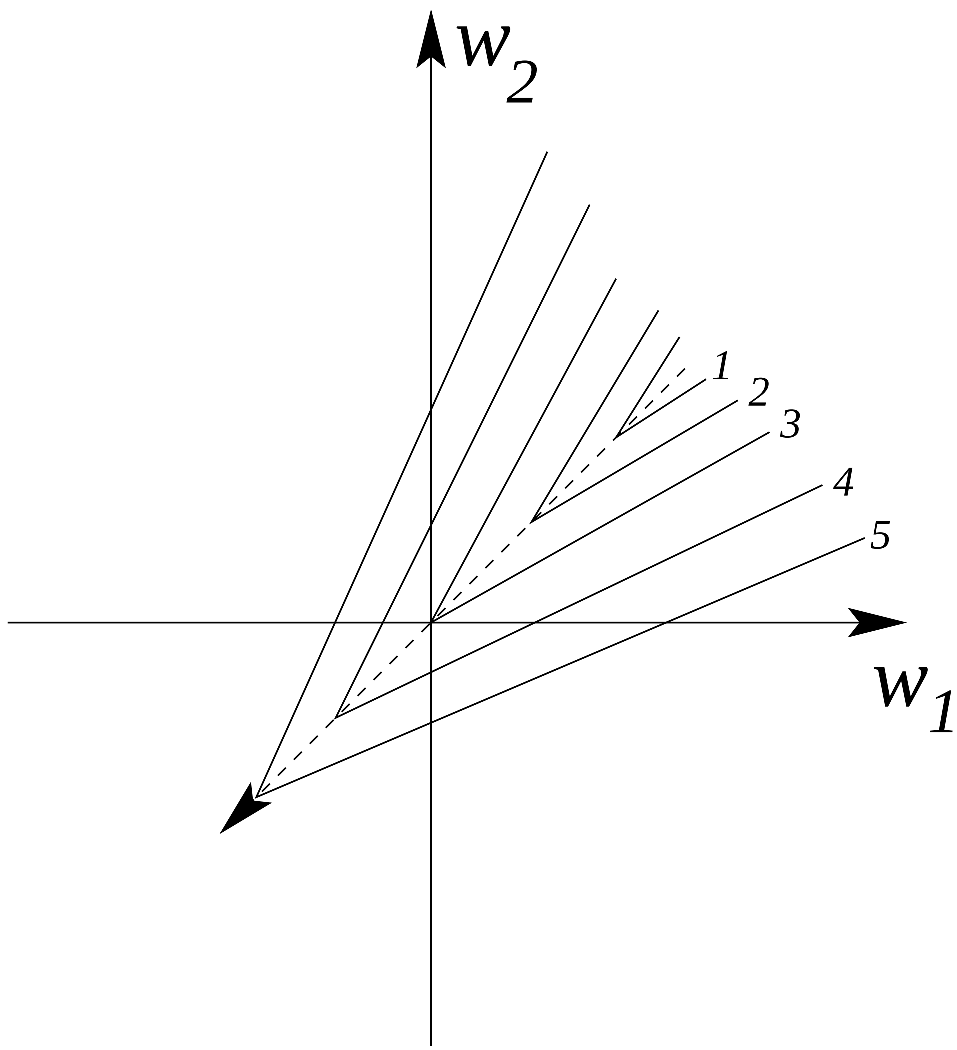


Non-smooth optimization

- **Simple techniques might not work!**
  - Gradient descent or coordinate descent
- **Special tools**
  - Subgradients or directional derivatives
- Typically slower than smooth optimization...
- ... except in some regularized problems

Counter-example

Coordinate descent for nonsmooth objectives



Regularized problems - Proximal methods

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

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

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

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

Cheap (and not dirty) algorithms for all losses

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

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

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

- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 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

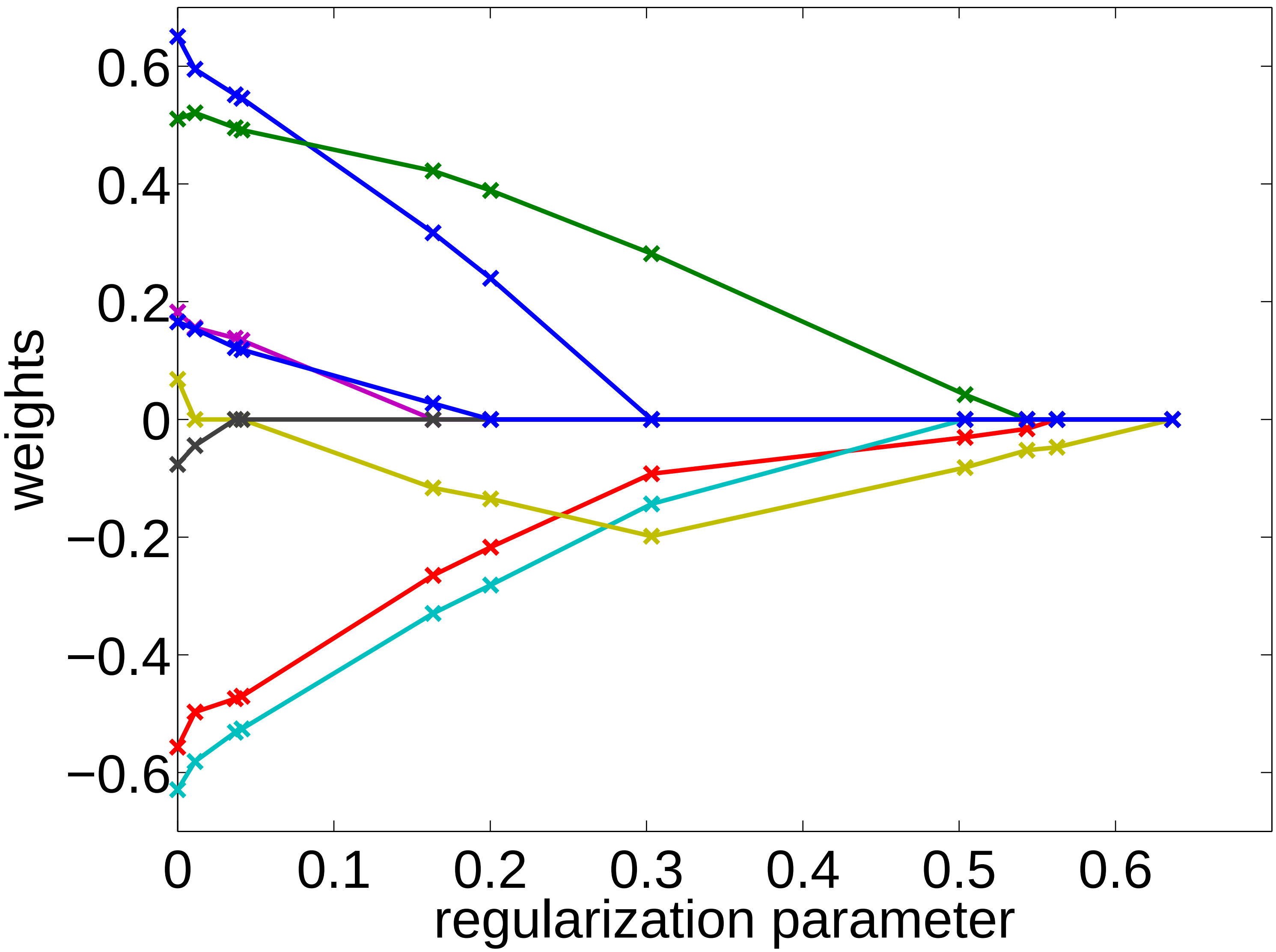
- **Proximal methods**

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

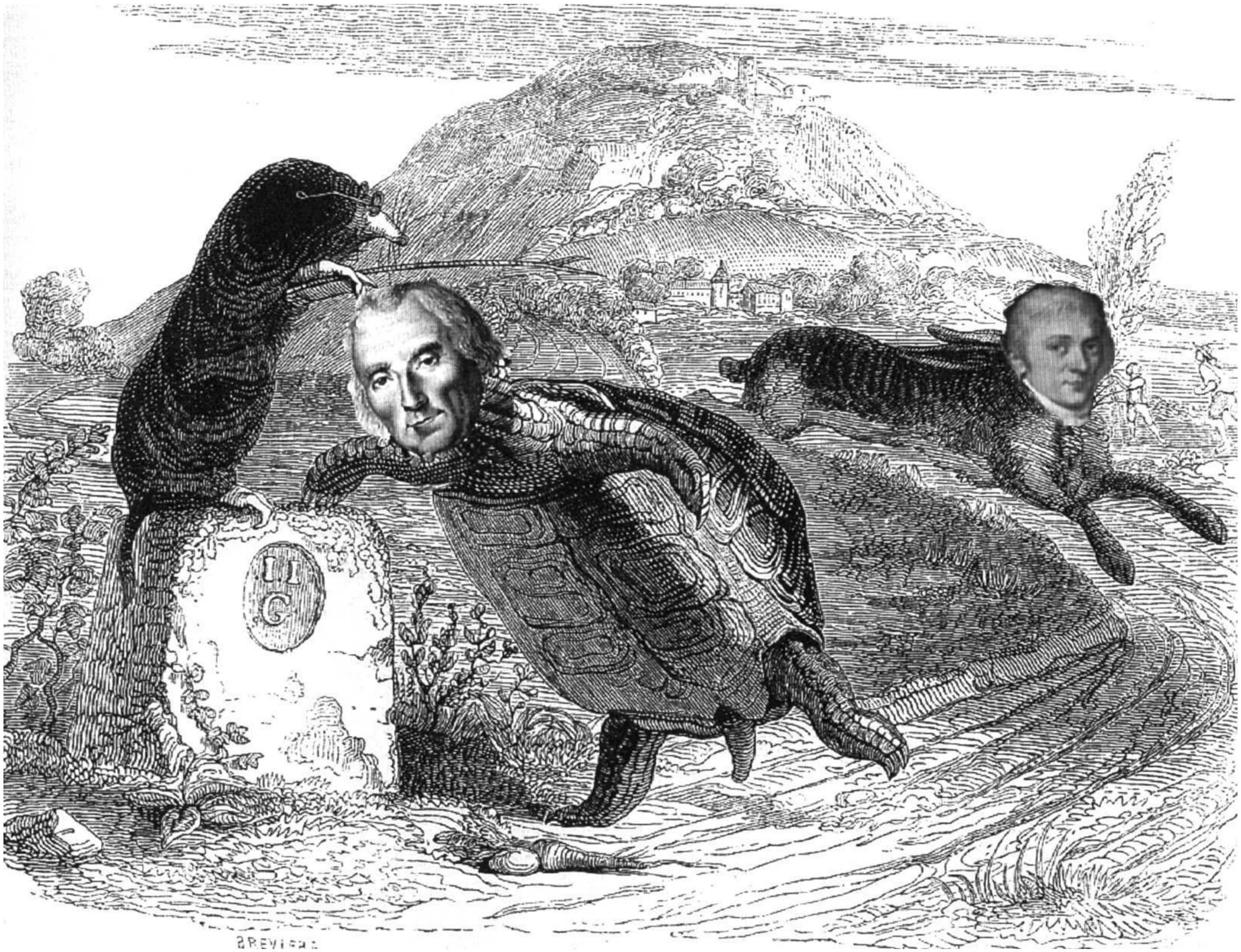
- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 2009b)
  - Notice that $\sum_{j=1}^{p} |w_j| = \min_{\eta \geqslant 0} \frac{1}{2} \sum_{j=1}^{p} \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to η (closed-form $\eta_j = |w_j|$) and w (weighted 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)

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/`

Lasso - Two main recent theoretical results

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

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)

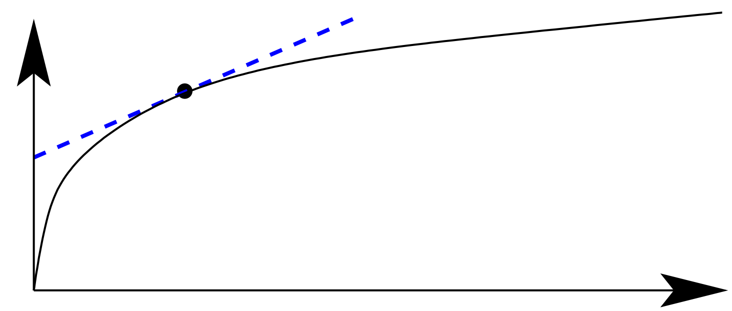
- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern
- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if
$$\boxed{\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1}$$
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 $\mathbf{J} = \operatorname{Supp}(\mathbf{w})$
- **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)

Lasso - Two main recent theoretical results

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

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

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

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
  - **and** input variables are not too correlated

High-dimensional inference (Lasso)

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

- Precise conditions on covariance matrix $\mathbf{Q} = \frac{1}{n} X^\top X$.
  - Mutual incoherence (Lounici, 2008)
  - Restricted eigenvalue conditions (Bickel et al., 2009)
  - Sparse eigenvalues (Meinshausen and Yu, 2008)
  - Null space property (Donoho and Tanner, 2005)

- Links with signal processing and compressed sensing (Candès and Wakin, 2008)

- **Slow rate if no assumptions**: $\sqrt{\frac{k \log p}{n}}$

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

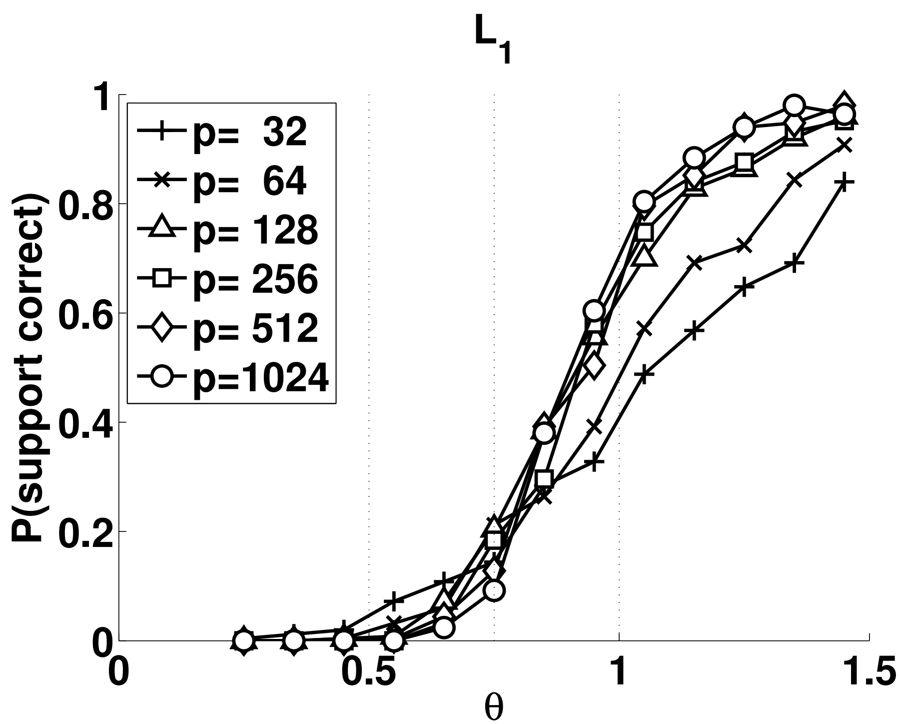
$$\text{estimation error} \qquad \|\hat{w} - \mathbf{w}\|_1 \leqslant \frac{16A}{\kappa^2(k)} \sigma k \sqrt{\frac{\log p}{n}}$$

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

- 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): $\boxed{\theta = \dfrac{n}{2k \log p} > 1}$



Sparse methods
Common extensions

- **Removing bias of the estimator**
  - Keep the active set, and perform unregularized restricted estimation (Candès and Tao, 2007)
  - Better theoretical bounds
  - Potential problems of robustness

- **Elastic net** (Zou and Hastie, 2005)
  - Replace $\lambda\|w\|_1$ by $\lambda\|w\|_1 + \varepsilon\|w\|_2^2$
  - Make the optimization strongly convex with unique solution
  - Better behavior with heavily correlated variables

Relevance of theoretical results

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

Alternative sparse methods
Greedy methods

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

Alternative sparse methods
Bayesian methods

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

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

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

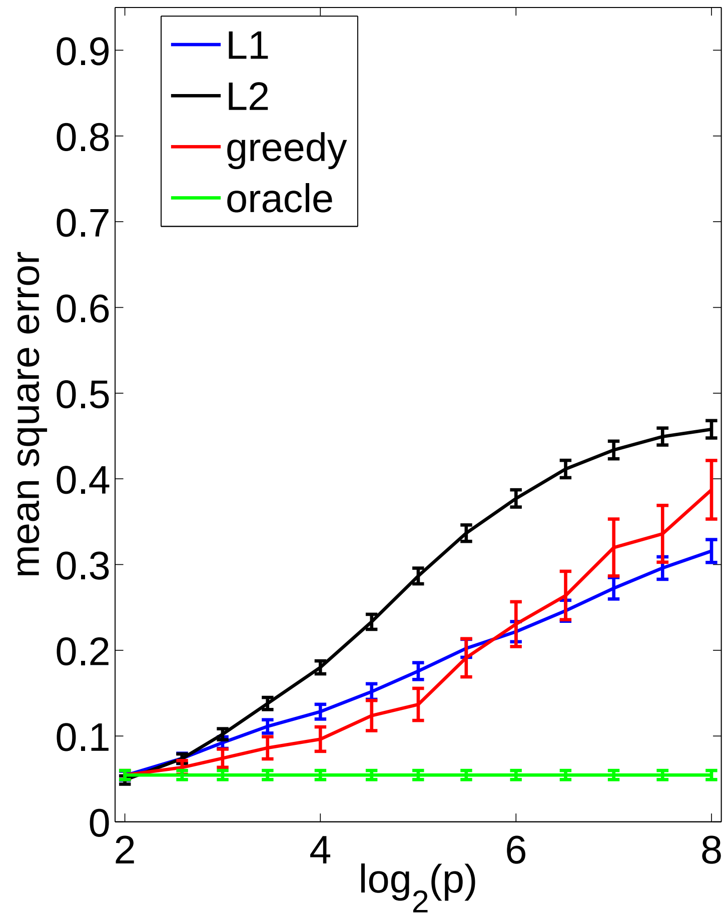
- Less theory than frequentist methods

Comparing Lasso and other strategies for linear regression

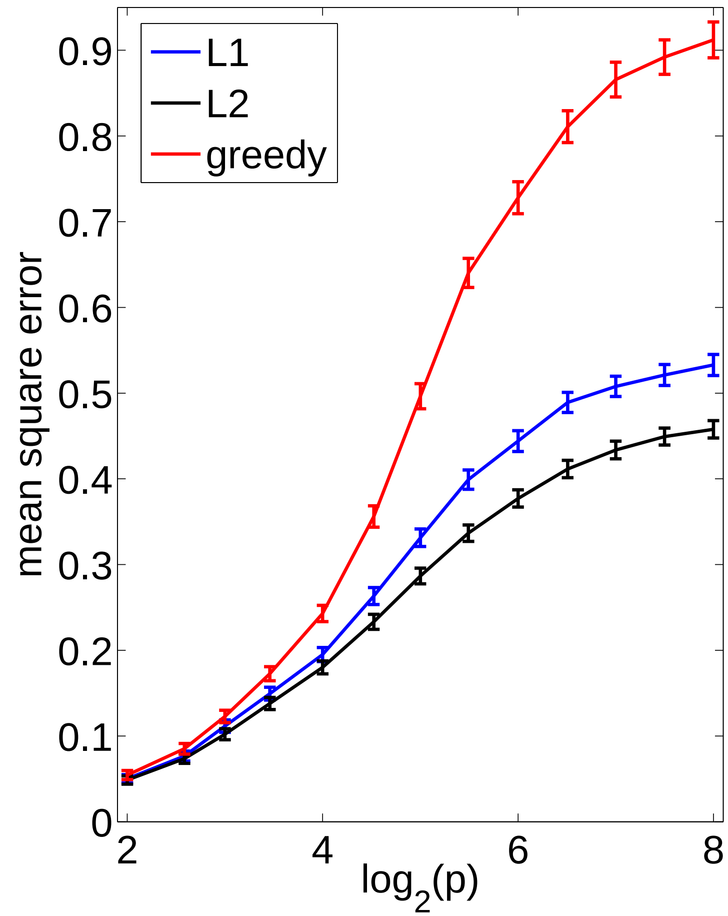
- Compared methods to reach the least-square solution
  - Ridge regression: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
  - Lasso: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$
  - Forward greedy:
    - Initialization with empty set
    - Sequentially add the variable that best reduces the square loss
- Each method builds a path of solutions from 0 to ordinary least-squares solution
- Regularization parameters selected on the test set

Simulation results

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



Sparse



Rotated (non sparse)

Going beyond the Lasso

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

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 $\sum_{j=1}^{p} \|w_j\|_2$

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

Going beyond the Lasso
Sparse methods on matrices

- **Learning problems on matrices**
  - Multi-task learning
  - Multi-category classification
  - Matrix completion
  - Image denoising
  - NMF, topic models, etc.

- **Matrix factorization**
  - Two types of sparsity (low-rank or dictionary learning)

Outline

- **Introduction: Sparse methods for machine learning**
  - **Short tutorial**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **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 $y = (y^1, \ldots, y^k) \in \mathbb{R}^{n \times k}$

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

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

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

  - 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 $x = (x^1, \dots, x^k) \in \mathbb{R}^{n \times k}$

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

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

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

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

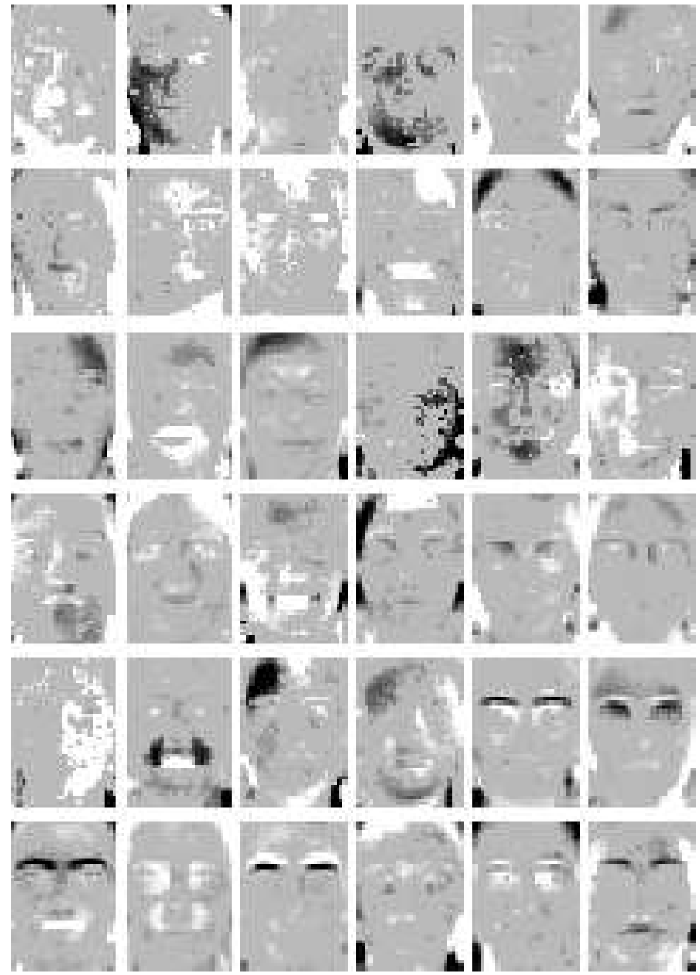
- **sparse PCA**: replace $\|d^j\|_2 \leqslant 1$ by $\Theta(d^j) \leqslant 1$

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., 2010)

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

raw data

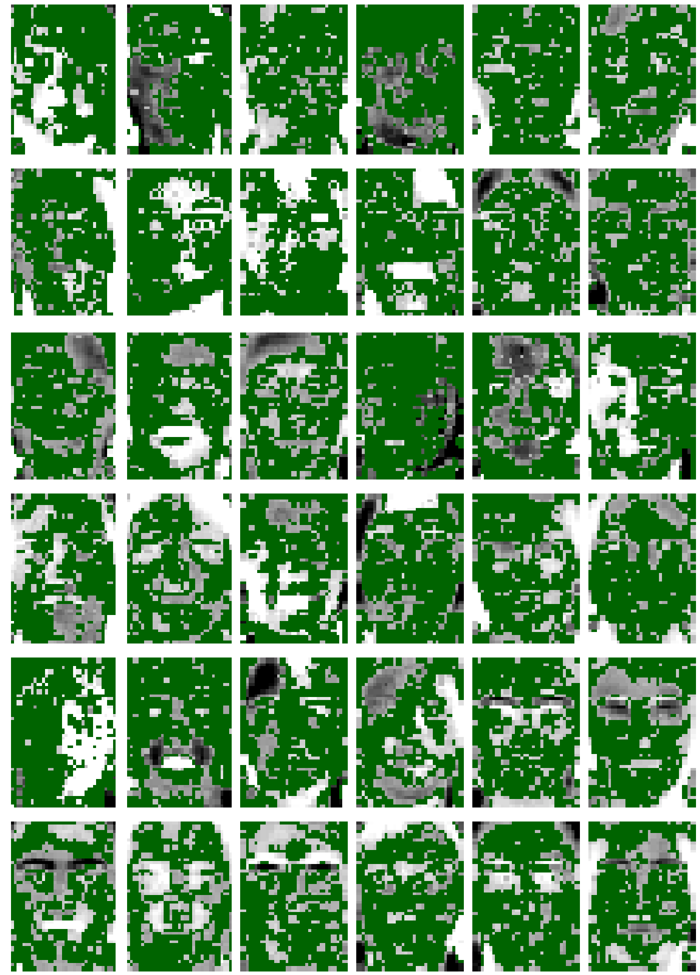


sparse PCA

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

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

raw data

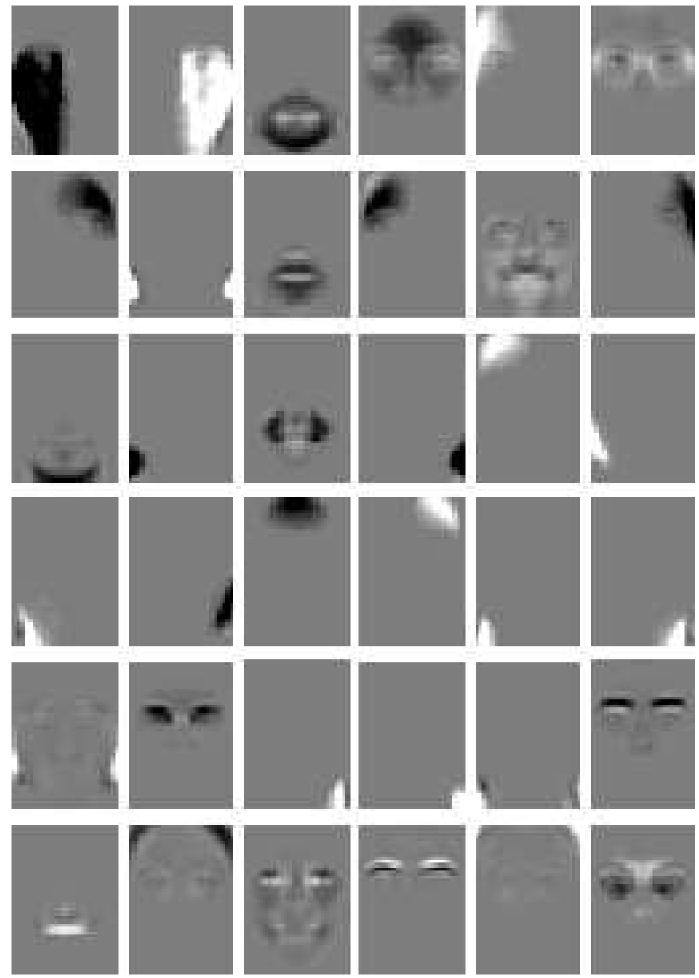


sparse PCA

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

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

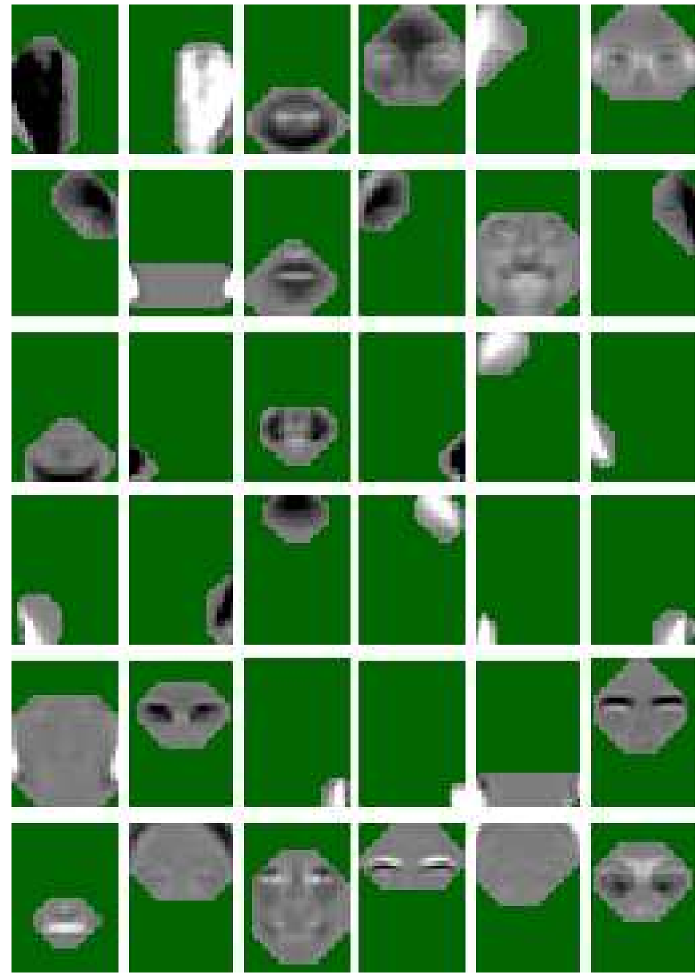
raw data



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., 2010)

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., 2010)
- **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., 2009c)
- **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., 2009b)
  - 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

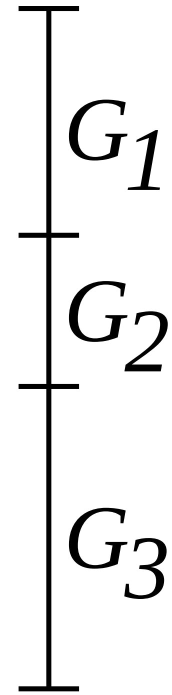
Outline

- **Introduction: Sparse methods for machine learning**
  - **Short tutorial**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **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

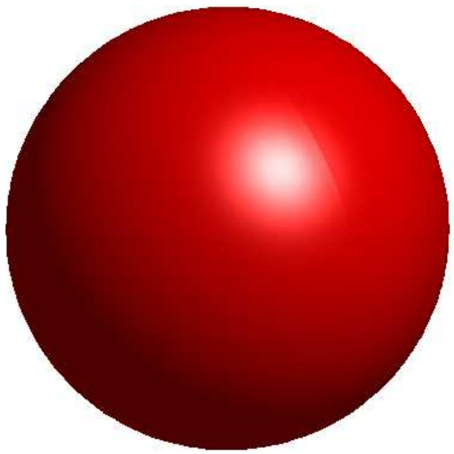
- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

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

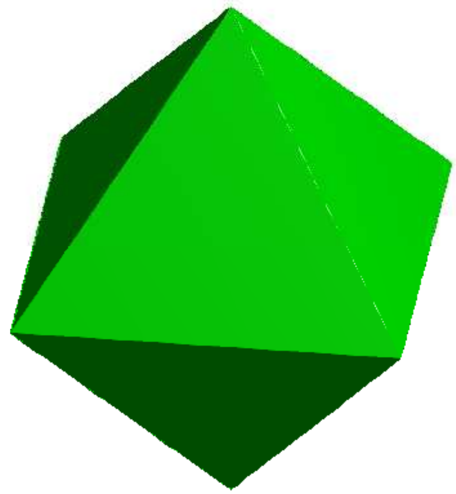


  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

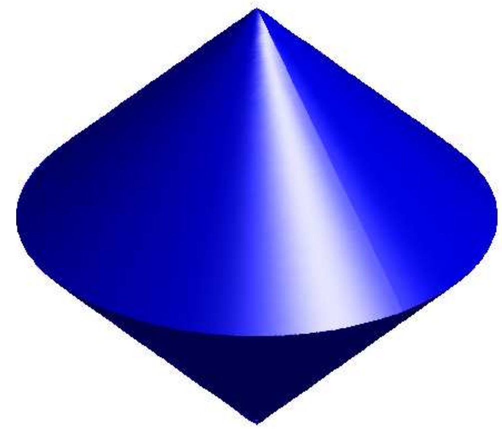
Unit norm balls
Geometric interpretation



$\|w\|_2$



$\|w\|_1$

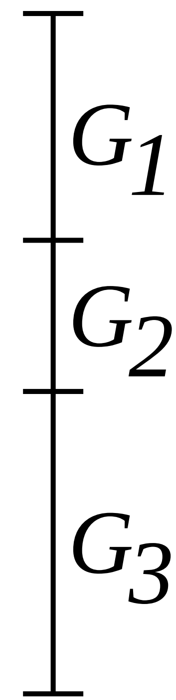


$\sqrt{w_1^2 + w_2^2} + |w_3|$

Sparsity-inducing norms

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

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

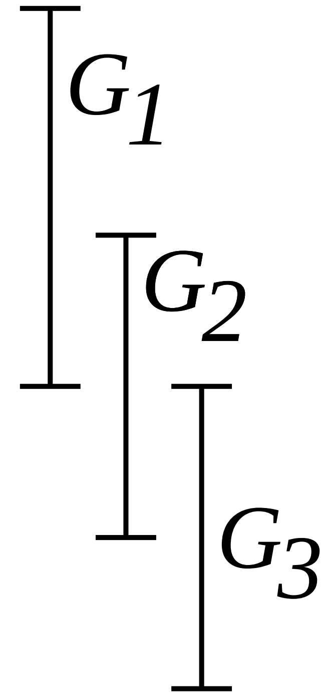


  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

- 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)

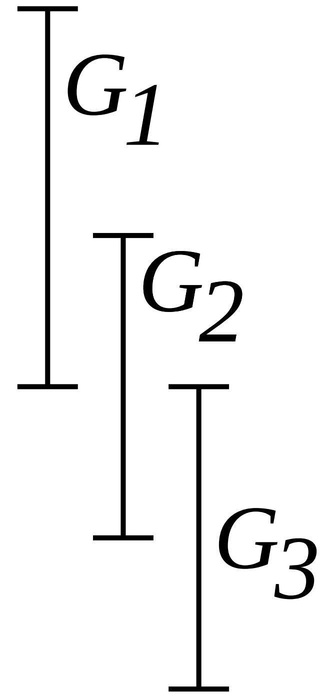


- When penalizing by the ℓ_1-ℓ_2 norm,

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

  – 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

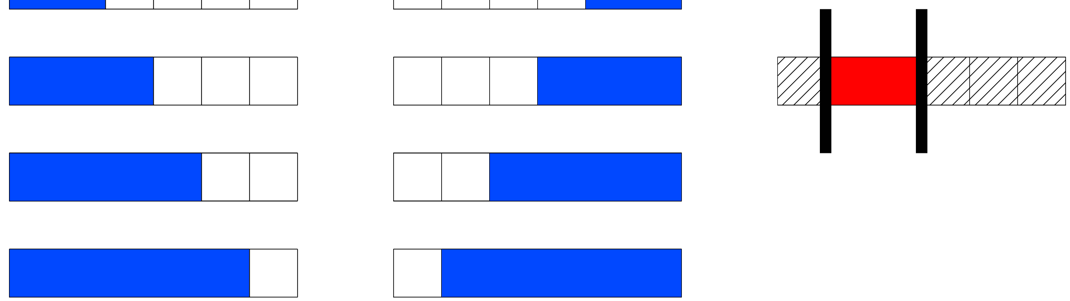
- The zero pattern of w is given by

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

- **Zero patterns are unions of groups**

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

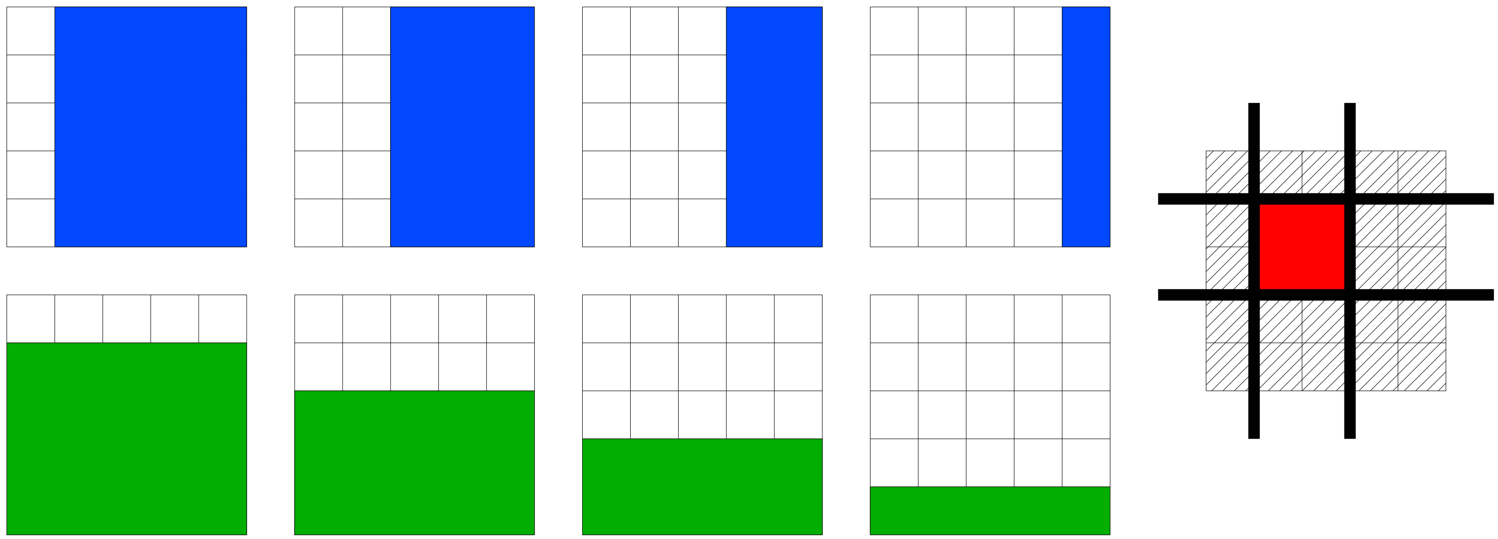
- Selection of contiguous patterns on a sequence, $p = 6$



  - $\mathbf{H}$ 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 $\mathbb{H}$

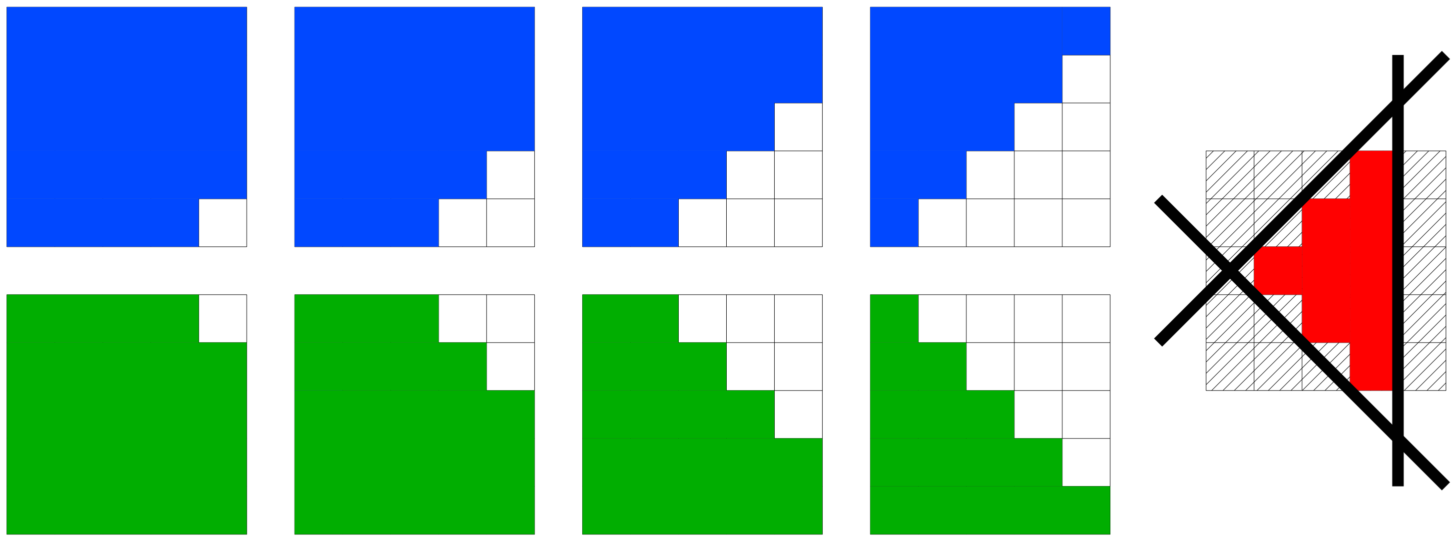
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ 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 $\mathbb{H}$

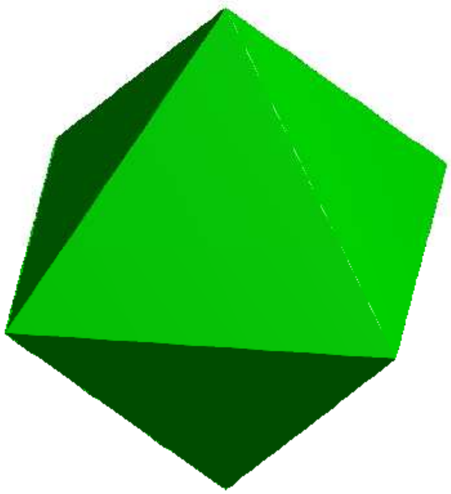
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



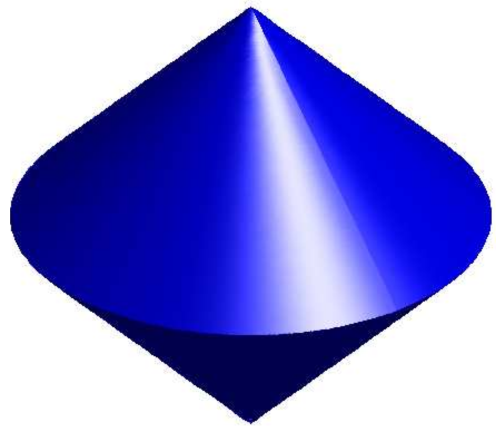
  - It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

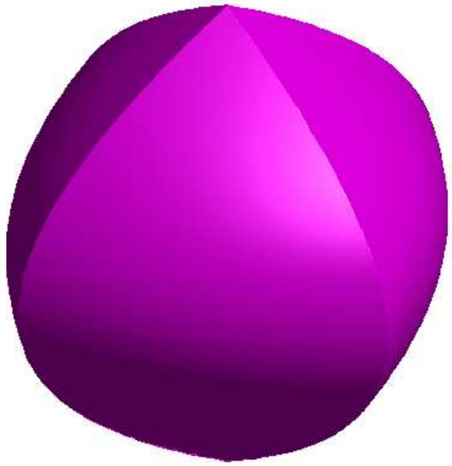
Geometric interpretation



$\|w\|_1$



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



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

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)

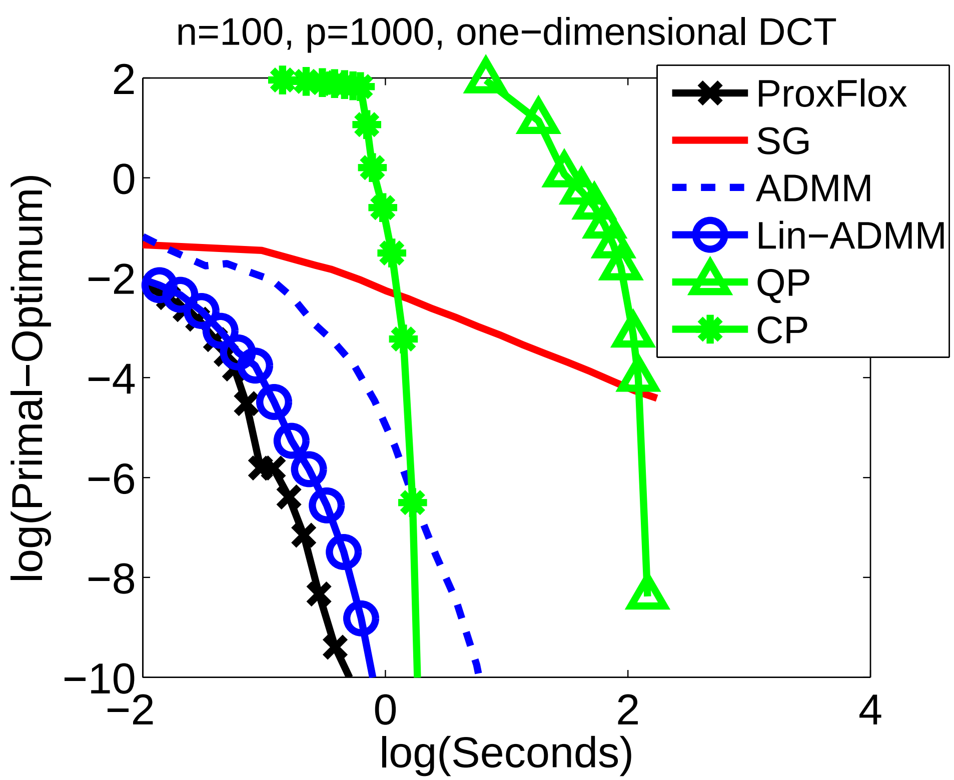
- Gradient descent as a **proximal method** (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{B} \nabla L(w_t)$

- 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)

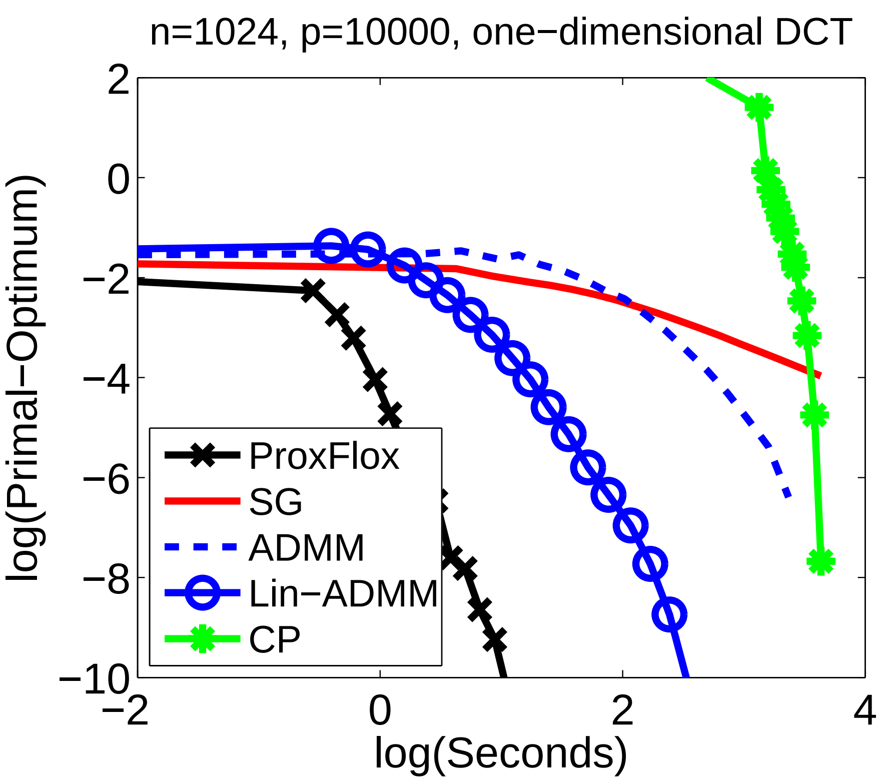
Comparison of optimization algorithms
(Mairal, Jenatton, Obozinski, and Bach, 2010)
Small scale

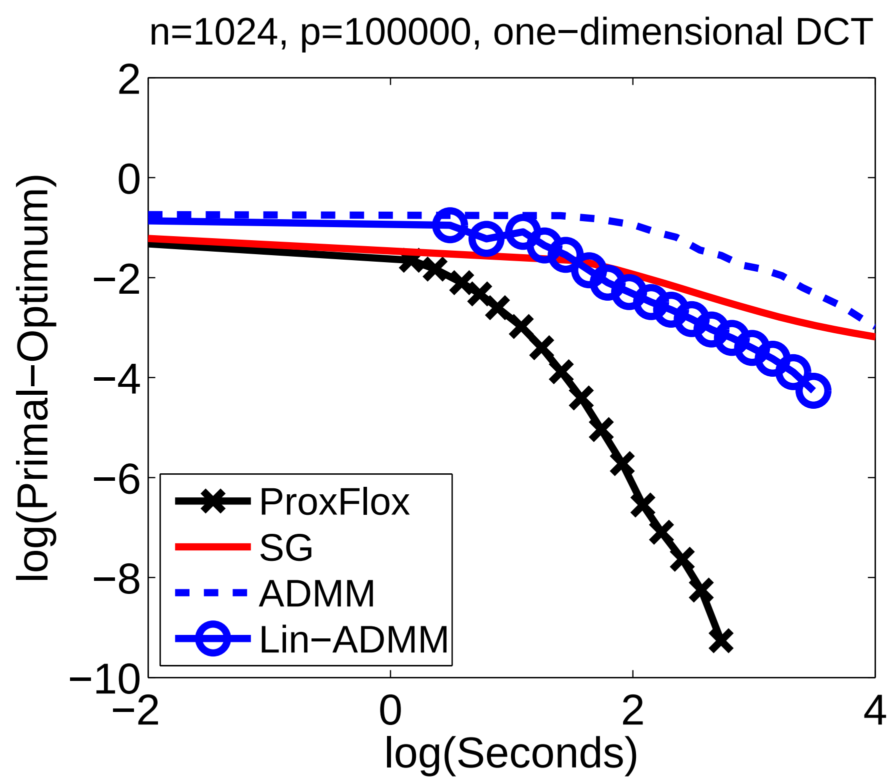
- Specific norms which can be implemented through network flows



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

- Specific norms which can be implemented through network flows





Approximate proximal methods (Schmidt, Le Roux, and Bach, 2011)

- Exact computation of proximal operator $\arg\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w-z\|_2^2 + \lambda\Omega(w)$
  - Closed form for ℓ_1-norm
  - Efficient for overlapping group norms (Jenatton et al., 2010; Mairal et al., 2010)

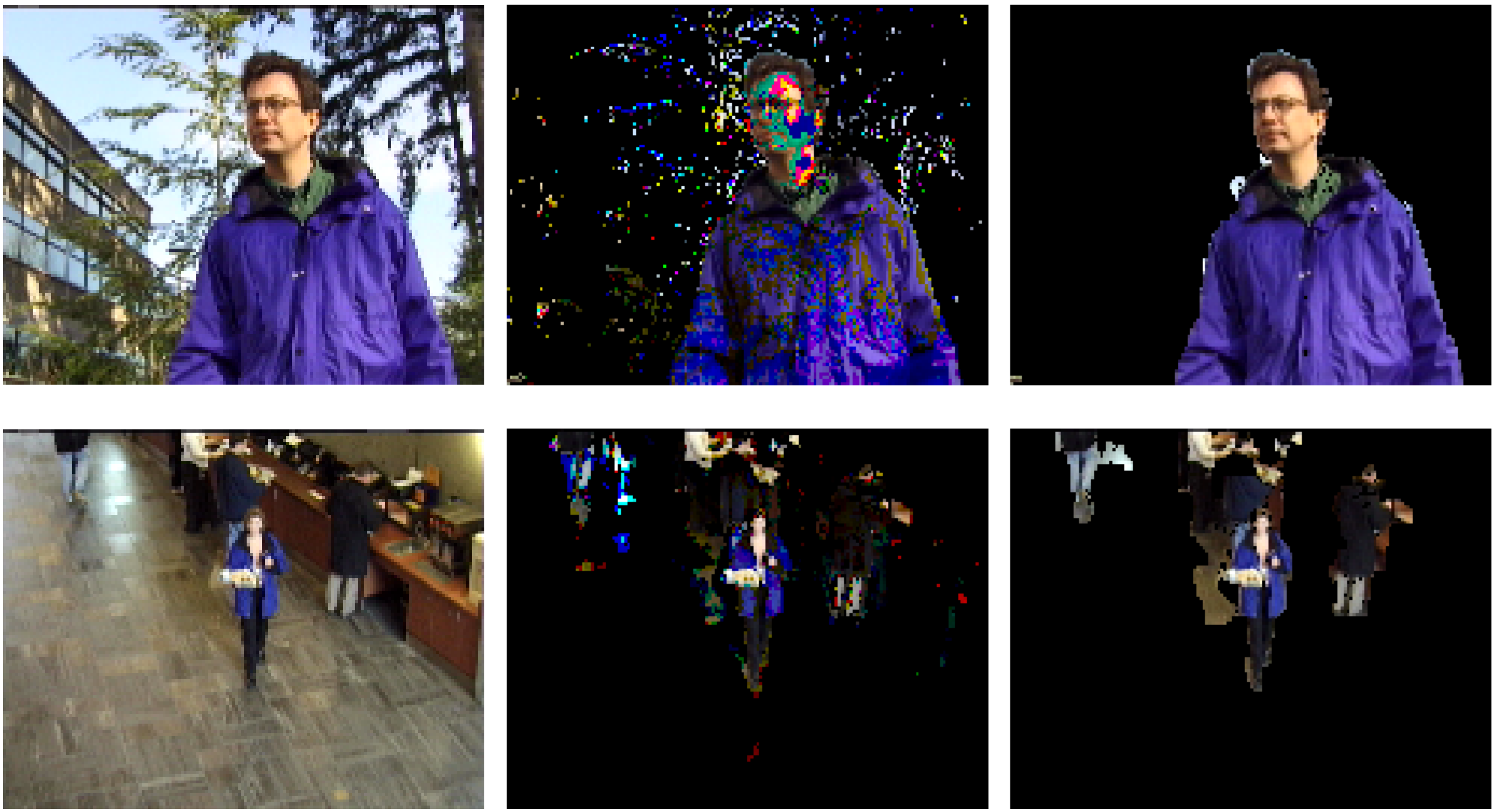
- Convergence rate: $O(1/t)$ and $O(1/t^2)$ (with acceleration)

- **Gradient or proximal operator may be only approximate**
  - Preserved convergence rate with appropriate control
  - Approximate gradient with non-random errors
  - Complex regularizers

Application to background subtraction

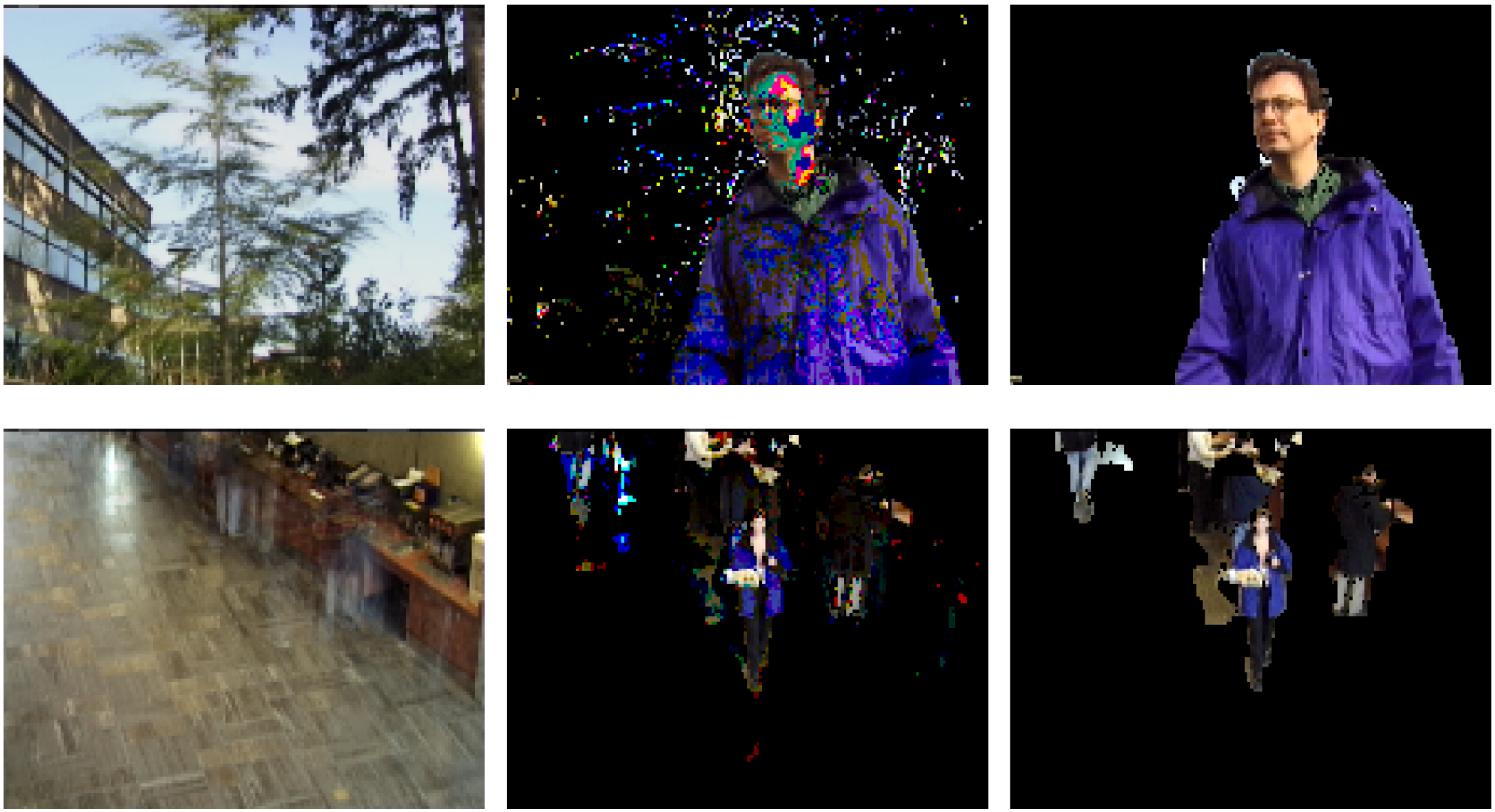
(Mairal, Jenatton, Obozinski, and Bach, 2010)

Input | ℓ_1-norm | Structured norm



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

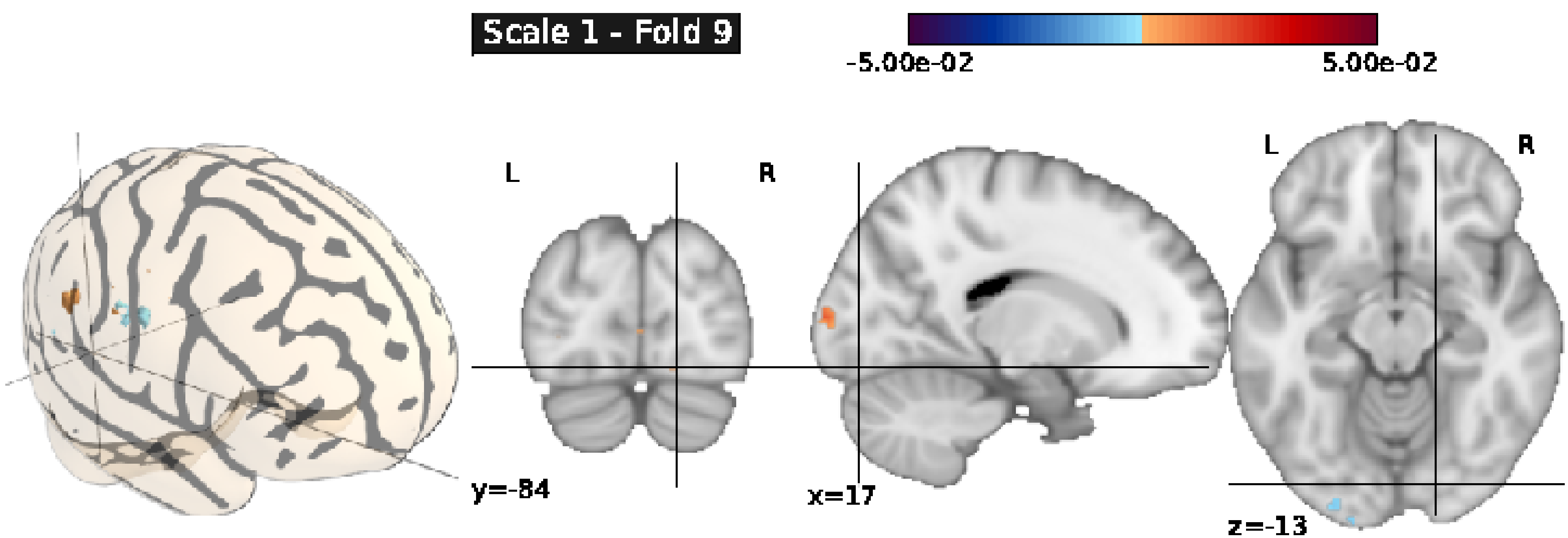
Background | ℓ_1-norm | Structured norm



Application to neuro-imaging

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

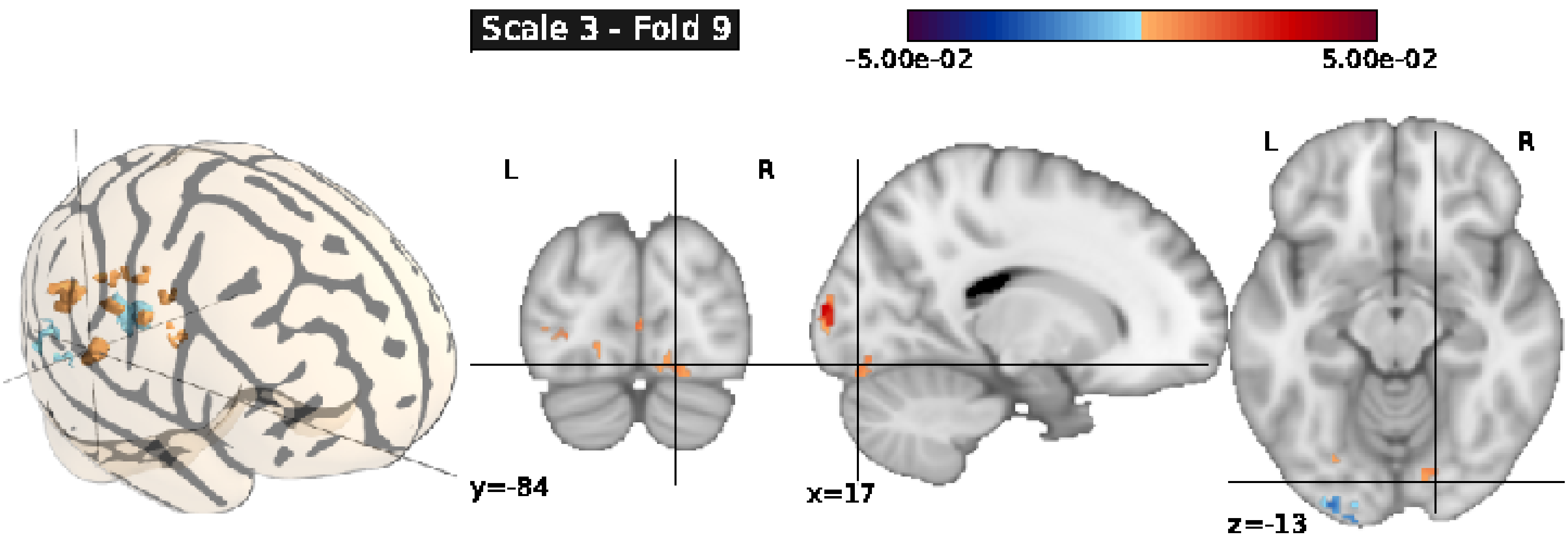
- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

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

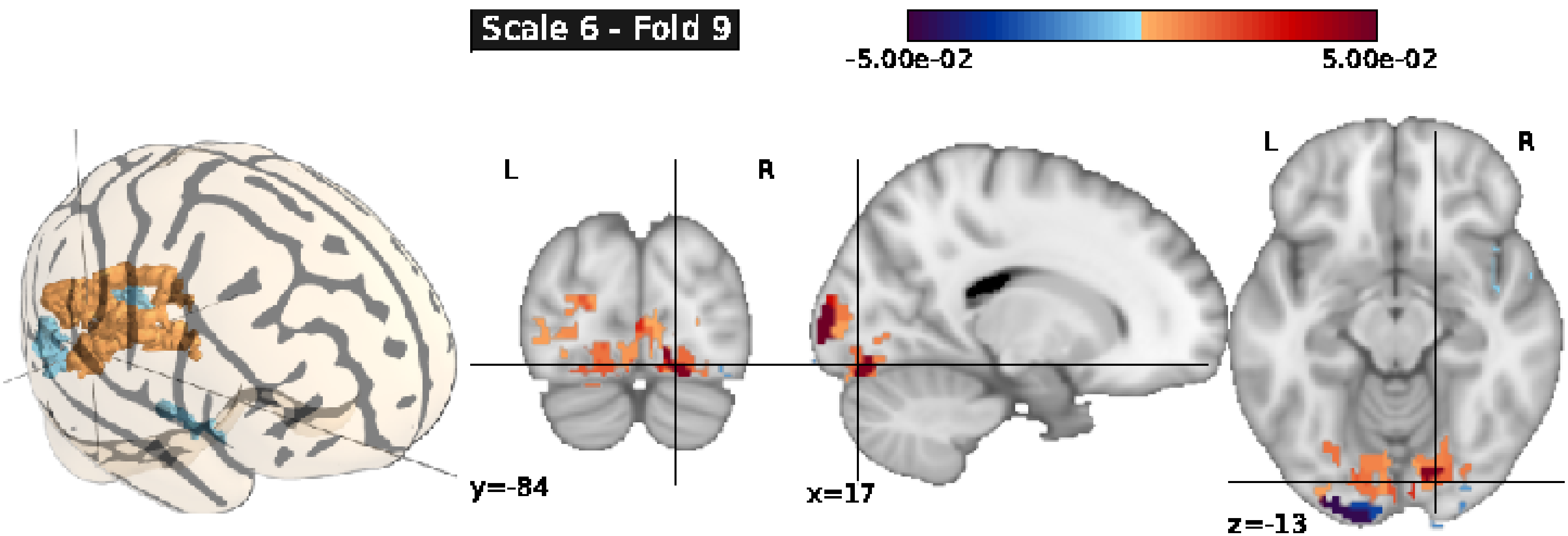
- "Brain reading": prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



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

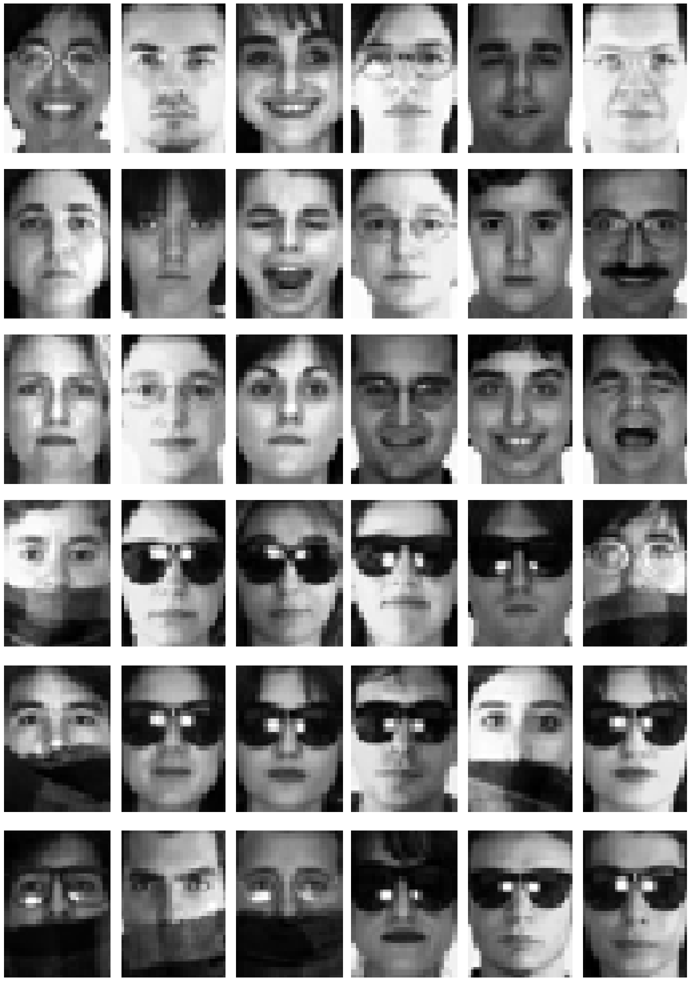


Sparse Structured PCA
(Jenatton, Obozinski, and Bach, 2009b)

- 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)

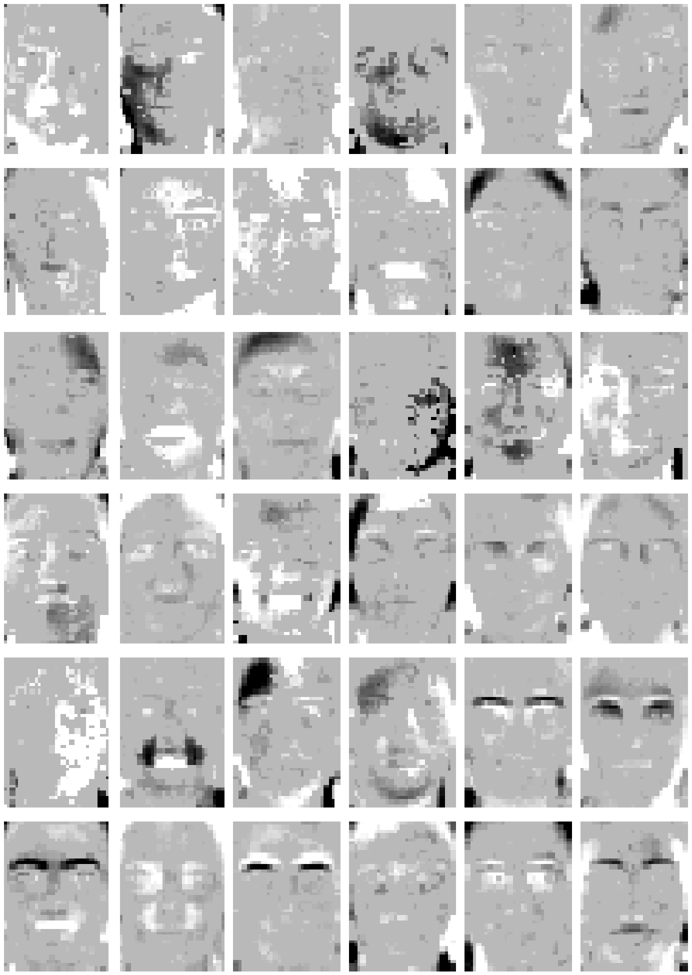


raw data

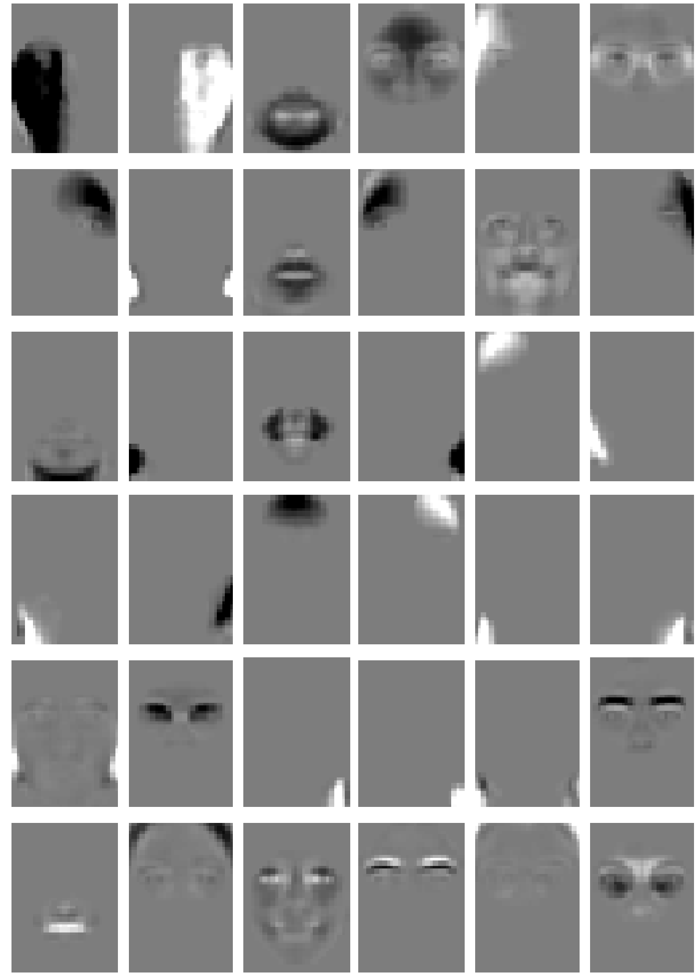
(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)



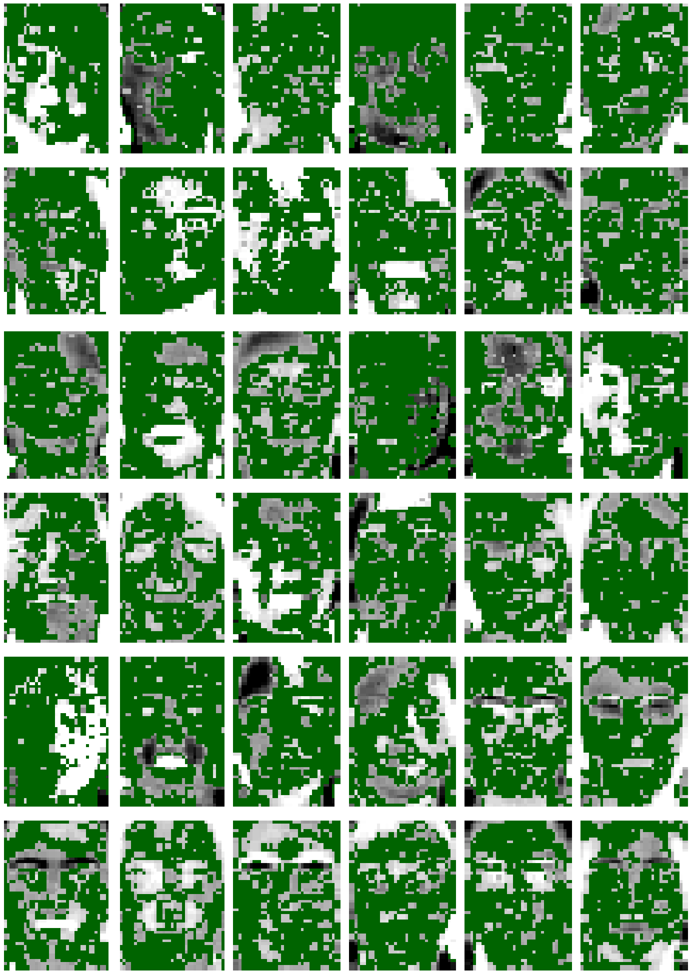
(unstructured) sparse PCA



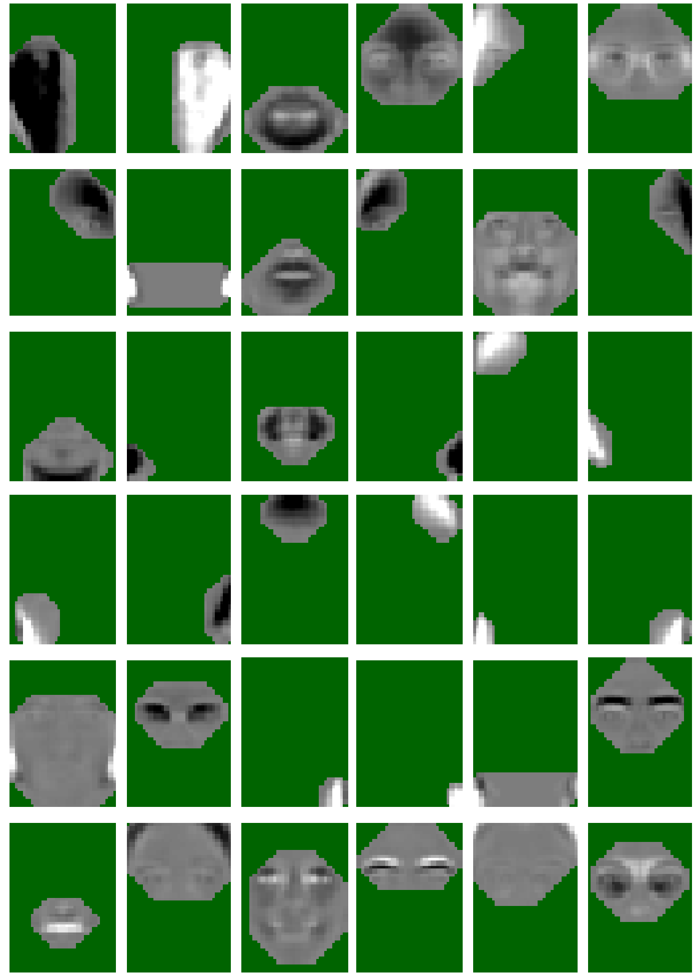
Structured sparse PCA

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

Application to face databases (2/3)



(unstructured) sparse PCA

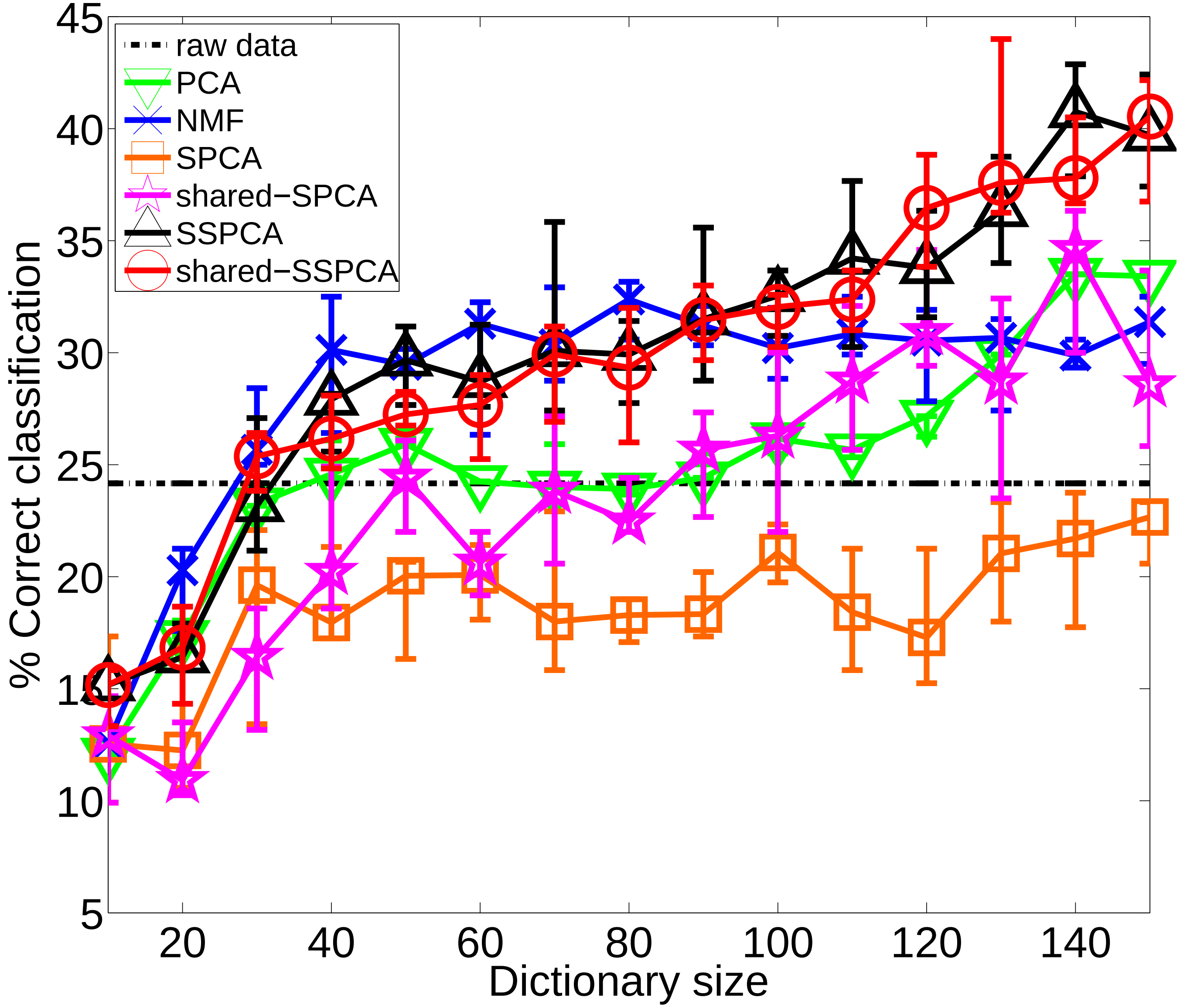


Structured sparse PCA

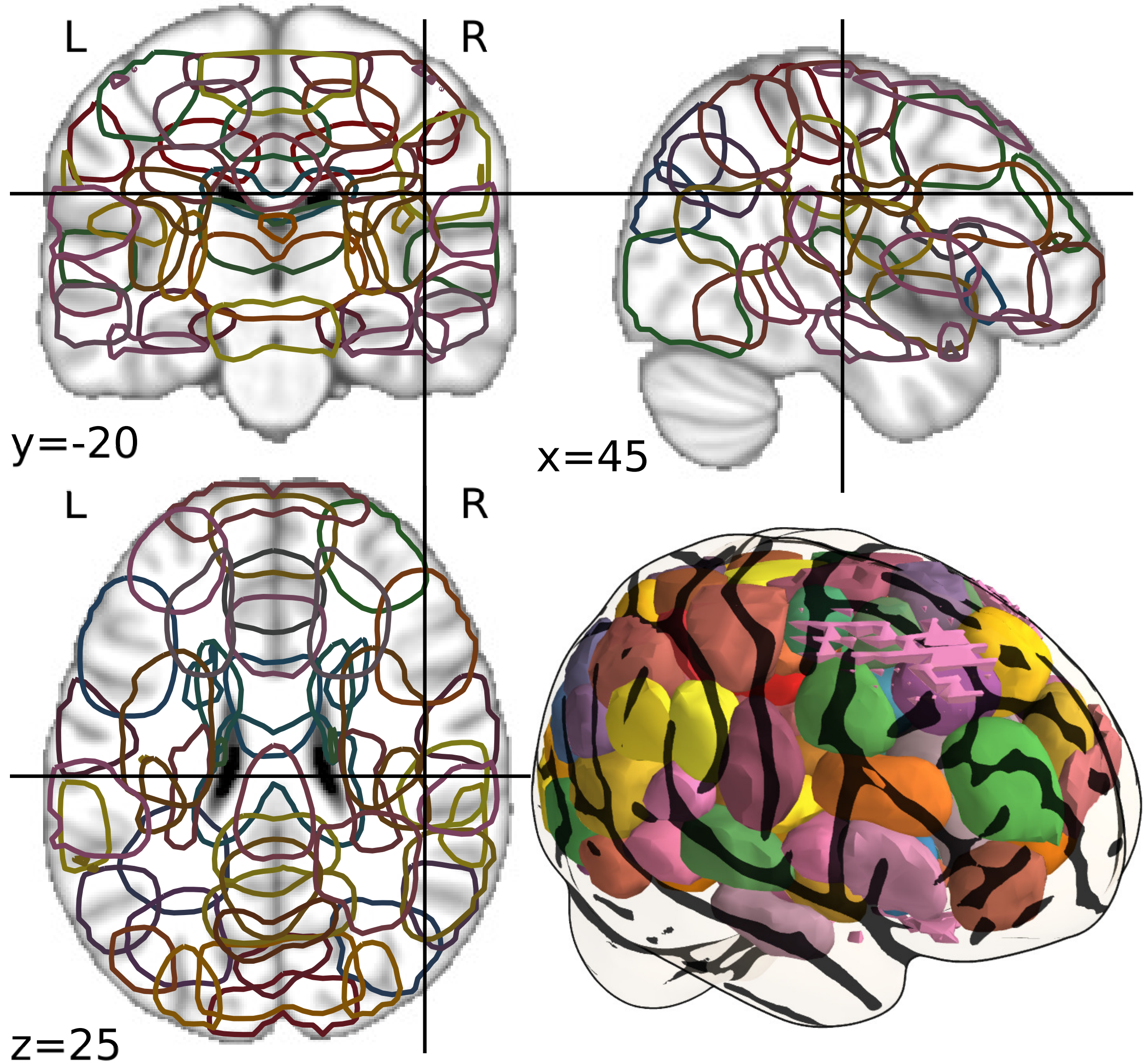
- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

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**):

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

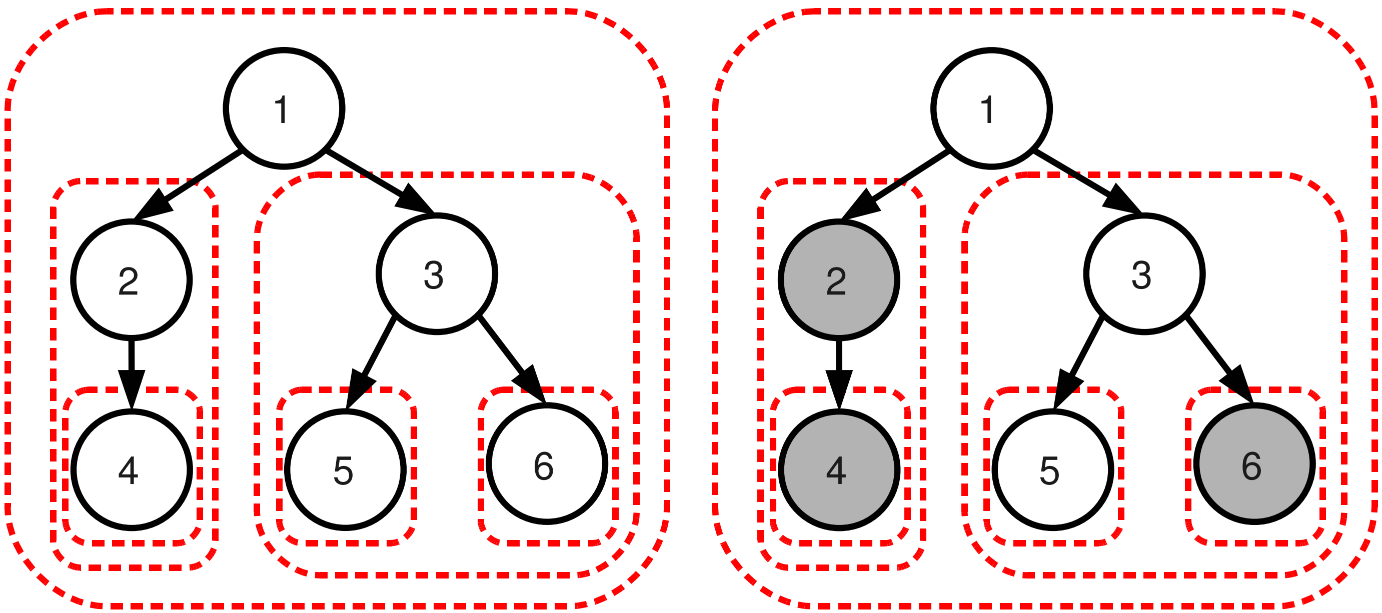
- Dictionary learning with **structured sparsity for codes** w:

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

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

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

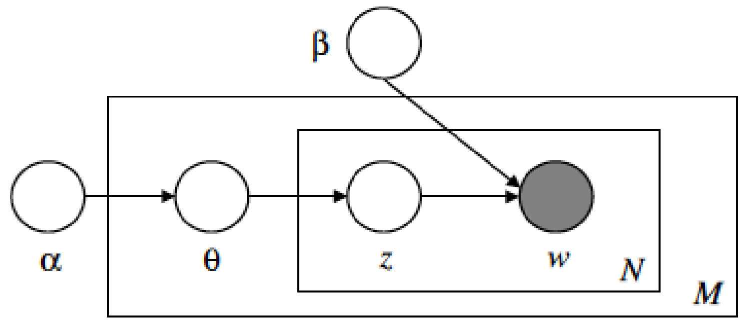


- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008c)

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., 2003a)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**

Topic models and matrix factorization

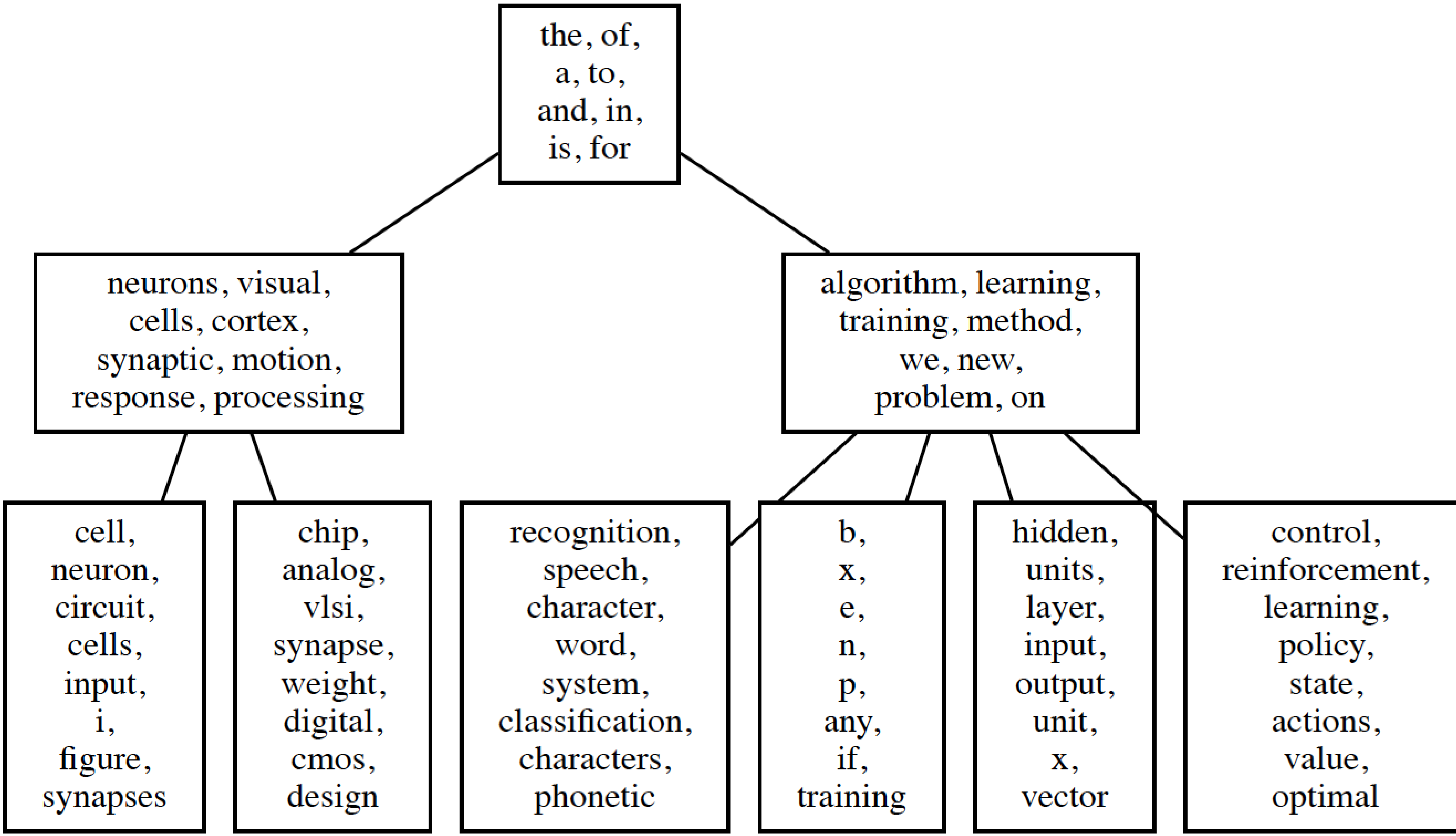


- **Latent Dirichlet allocation** (Blei et al., 2003b)
  - For a document, sample $\theta \in \mathbb{R}^k$ from a Dirichlet(α)
  - For the n-th word of the same document,
    * sample a topic z_n from a multinomial with parameter θ
    * sample a word w_n from a multinomial with parameter $\beta(z_n, :)$

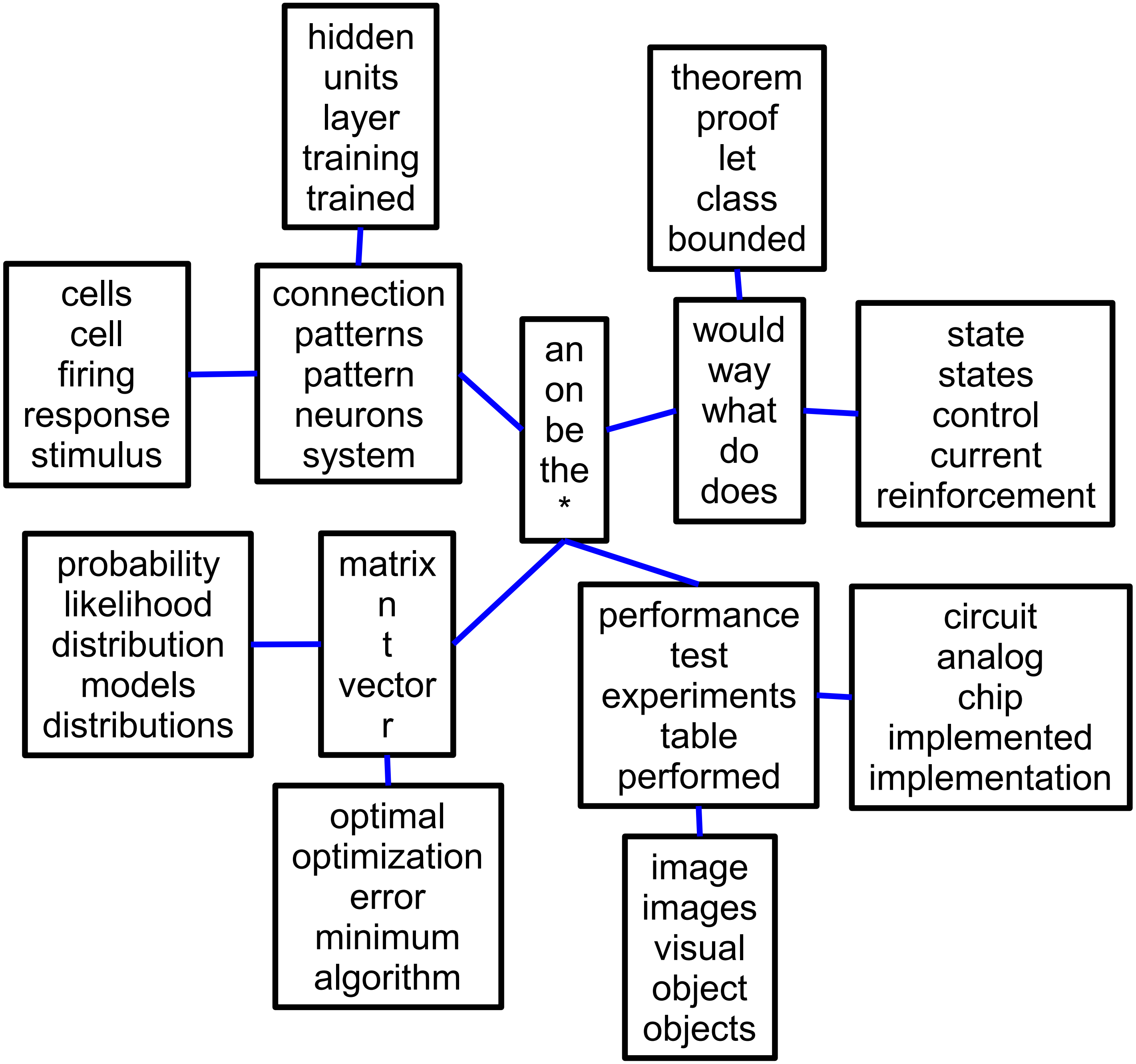
- **Interpretation as multinomial PCA** (Buntine and Perttu, 2003)
  - Marginalizing over topic z_n, given θ, each word w_n is selected from a multinomial with parameter $\sum_{z=1}^{k} \theta_z \beta(z, :) = \beta^\top \theta$
  - Row of β = dictionary elements, θ code for a document

Modelling of text corpora - Dictionary tree

Probabilistic topic models (Blei et al., 2004)



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., 2009c)
- **Optimization and local minima**

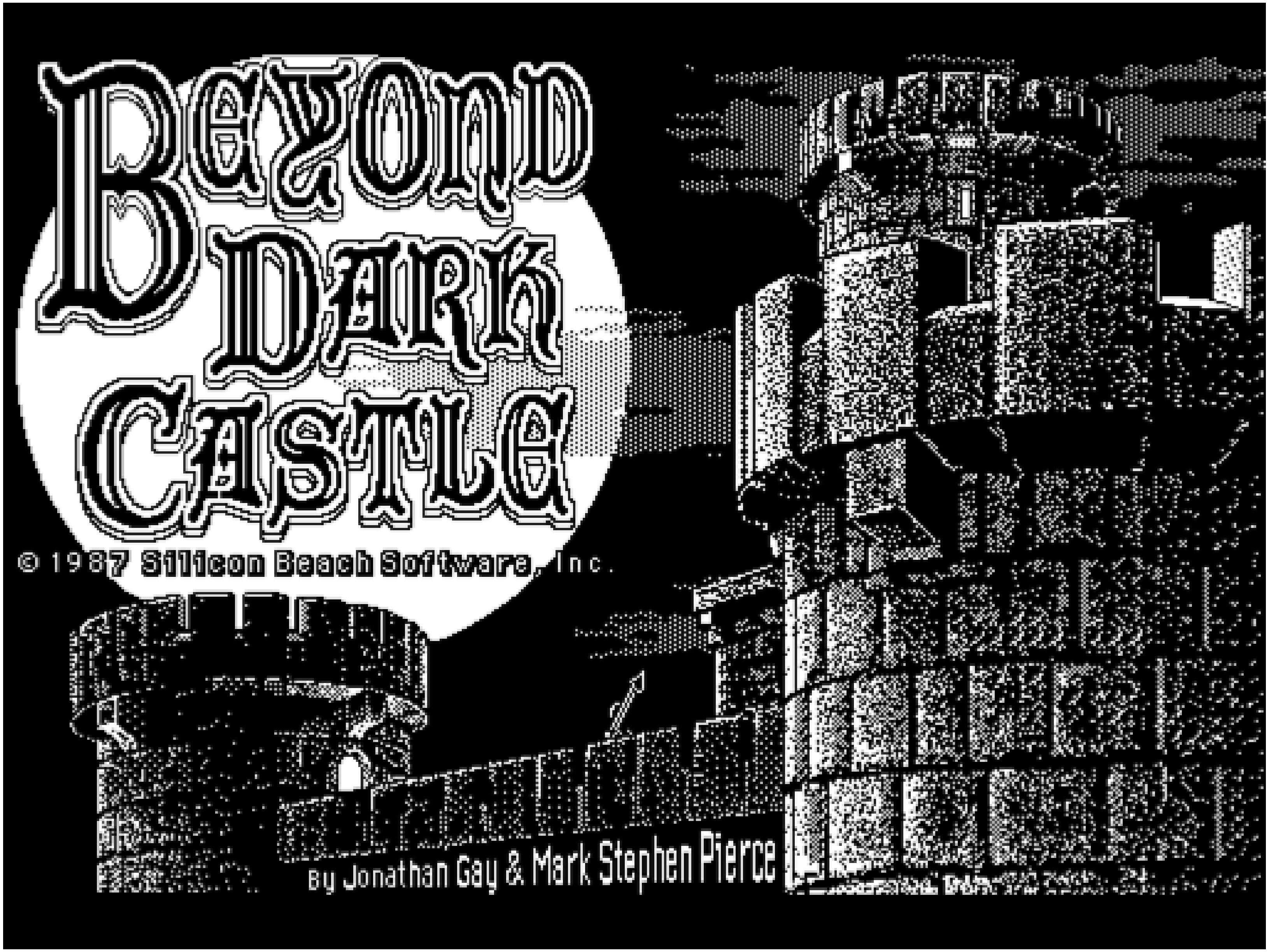
Digital zooming (Couzinie-Devy et al., 2011)

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

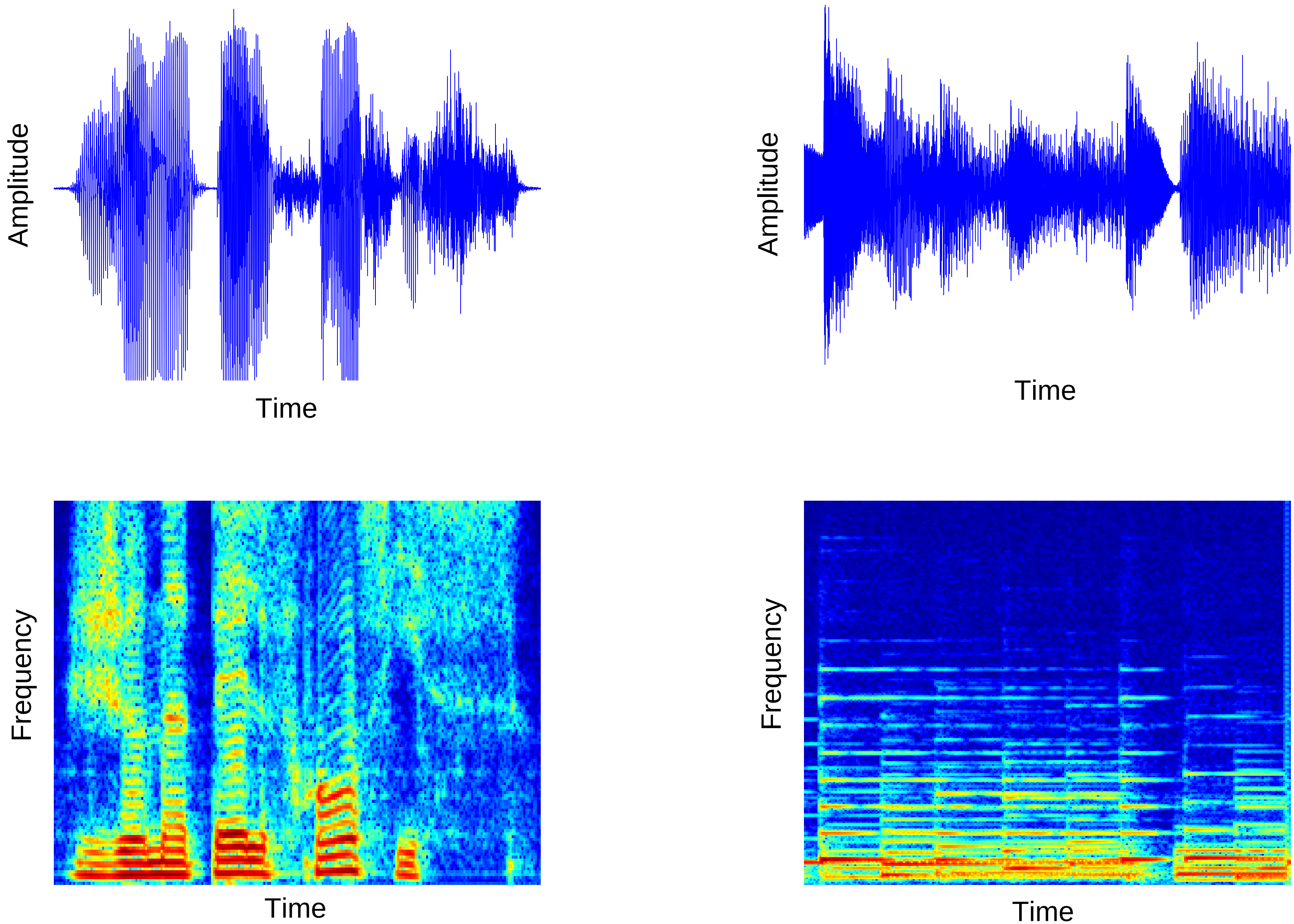


Ongoing Work - Inverse half-toning

Ongoing Work - Inverse half-toning

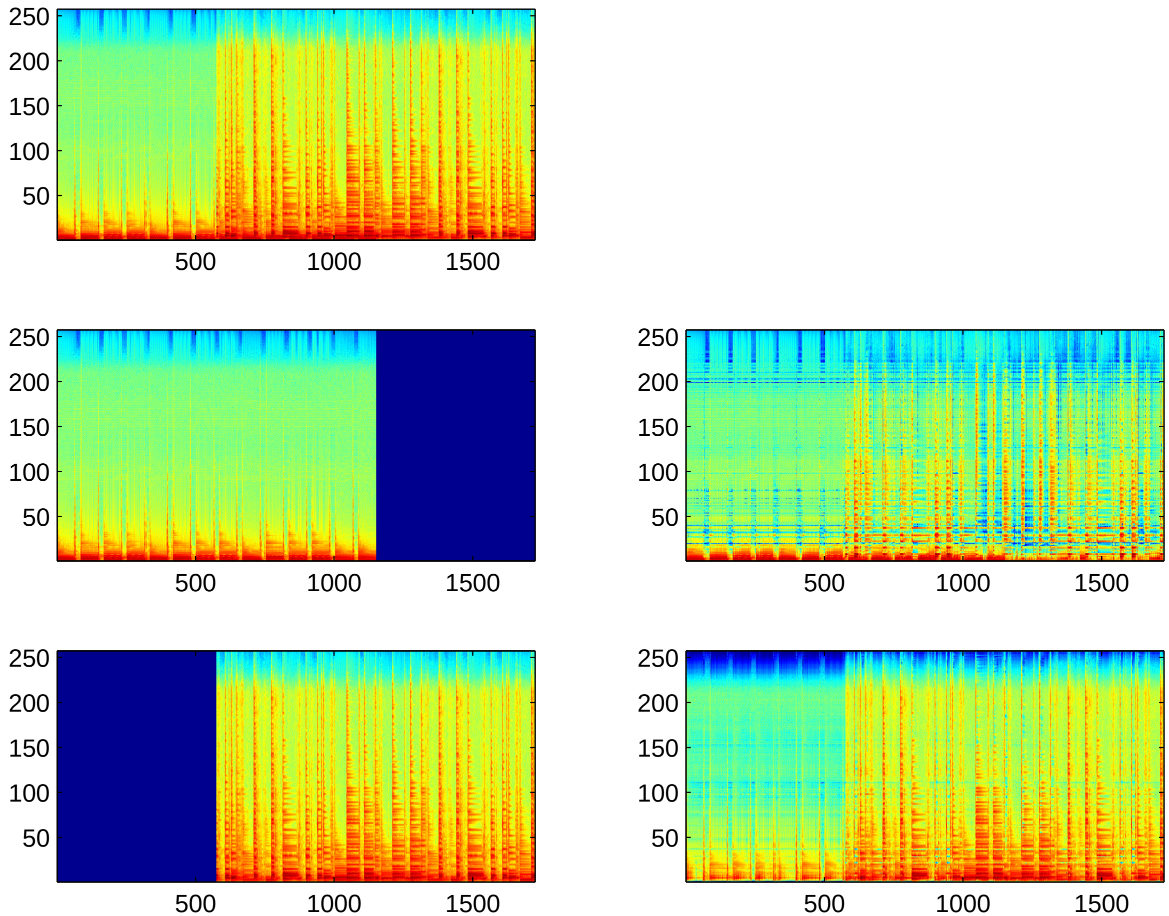
Ongoing Work - Inverse half-toning

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.

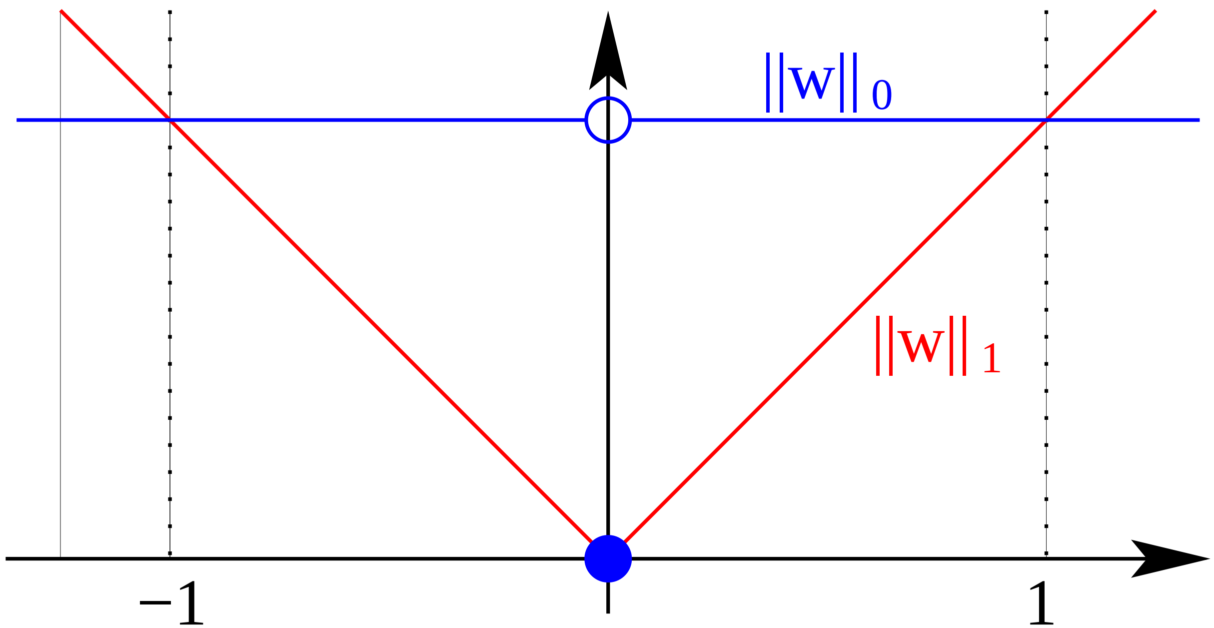


Outline

- **Introduction: Sparse methods for machine learning**
  - **Short tutorial**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **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 $\Theta(w) = F(\mathrm{Supp}(w))$: 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, 2010)

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

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

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

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ 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

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ 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
- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ 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:

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

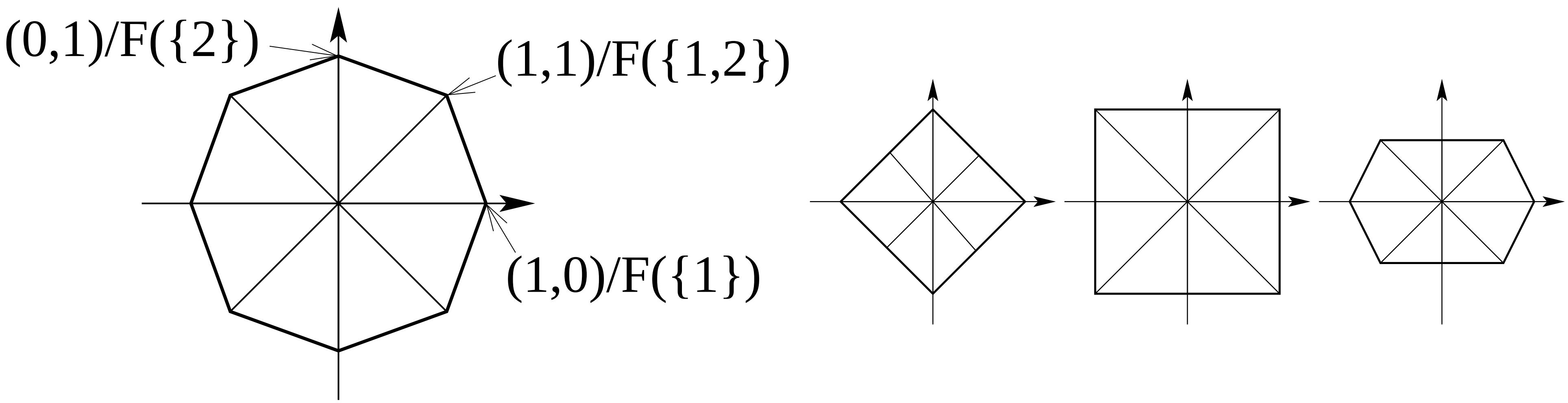
  - If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  - f is piecewise affine and positively homogeneous
- F is submodular if and only if f is convex (Lovász, 1982)
  - Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V

Submodular functions and structured sparsity

- Let $F : 2^V \to \mathbb{R}$ 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

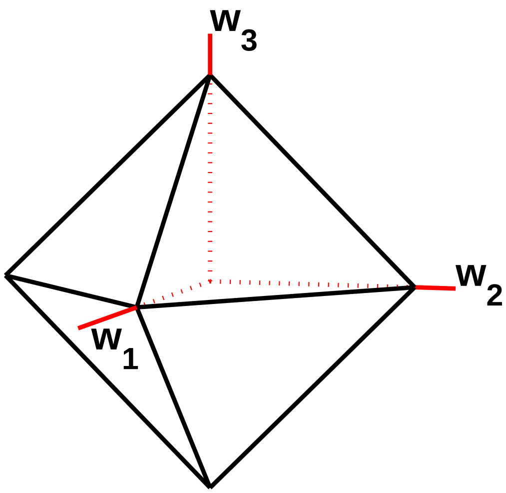
Submodular functions and structured sparsity

- 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

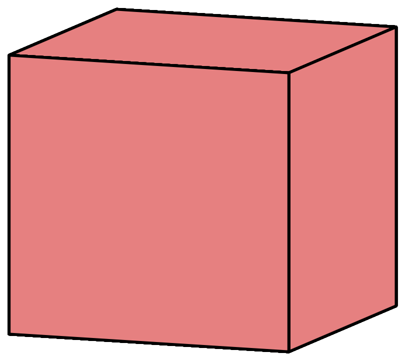


  - A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  - With probability one, stable sets are the only allowed active sets

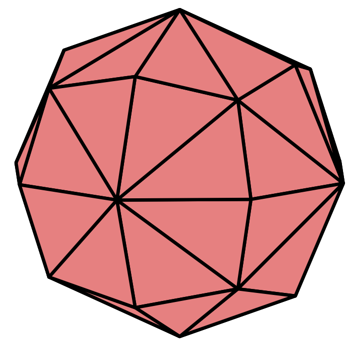
Polyhedral unit balls



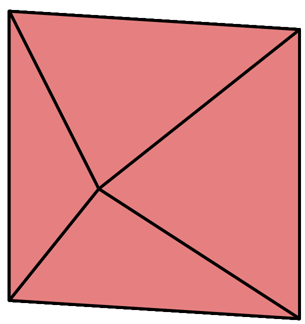
$F(A) = |A|$
$\Omega(w) = \|w\|_1$



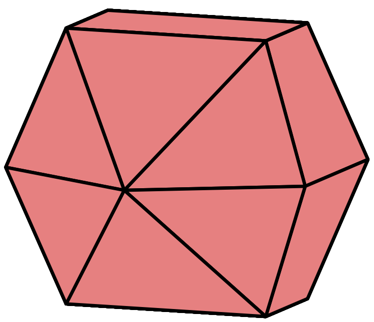
$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



$F(A) = |A|^{1/2}$
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)

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

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

- **On-going/related work on structured sparsity**
  - Norm design beyond submodular functions
  - Complementary approach of Jacob, Obozinski, and Vert (2009)
  - Theoretical analysis of dictionary learning (Jenatton, Bach and Gribonval, 2011)
  - Achieving $\log p = O(n)$ algorithmically (Bach, 2008c)

References

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

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

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

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

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

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

F. Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010.

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

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. 2011.

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

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.

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 2003a.

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.

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

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.

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

W. Buntine and S. Perttu. Is multinomial PCA multi-faceted clustering or dimensionality reduction. In

International Workshop on Artificial Intelligence and Statistics (AISTATS), 2003.

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.

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.

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

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

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

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.

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.

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

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.

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

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, 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, 2009b.

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

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

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. Stability selection. Technical report, arXiv: 0809.2932, 2008.

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

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

Y. Nesterov. 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.

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.

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.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Arxiv preprint arXiv:1109.2415*, 2011.

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

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

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

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.

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.