[80240603 Advanced Machine Learning, Fall, 2012]

Sparse Learning

Jun Zhu
dcszj@mail.tsinghua.edu.cn
State Key Lab of Intelligent Technology & Systems
Tsinghua University

December 13, 2012

Two Important Aspects

- Model goodness:
  - Often defined in terms of prediction accuracy
- Model parsimony:
  - Simpler models are preferred for the sake of scientific insight into the x – y relationship

Example – diabetes study

- 442 diabetes patients were measured on 10 baseline variables; a prediction model was desired for the response variable, a measure of disease progression one year after baseline

	AGE	SEX	BMI	BP	Serum measurements						Response
Patient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

- **Two hopes**:
  - Accurate prediction
  - Understand important factors

Supervised Learning and Regularization

- **Data**

$$x_i \in \mathcal{X}, \;\; y_i \in \mathcal{Y}, \;\; i = 1, 2, \ldots, n$$

- Minimize with respect to function $f : \mathcal{X} \to \mathcal{Y}$

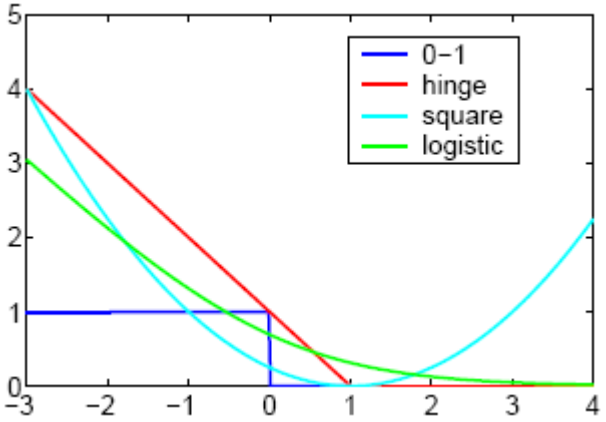
$$\sum_n \ell(f(x_i), y_i) \quad + \quad \frac{\lambda}{2}\|f\|^2$$

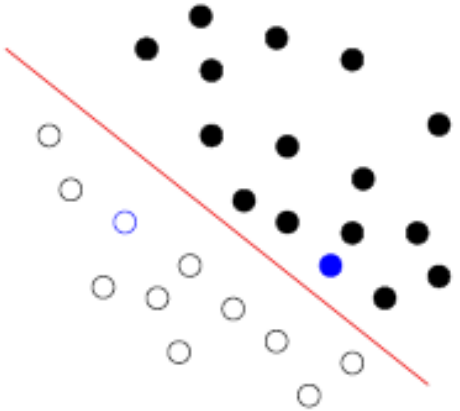
Error on data + Regularization

- Two theoretical/algorithmic issues
  - Loss
  - Function space/norm

Usual Losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = f(x)$
  - quadratic cost is $\ell(y, f) = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - f)^2$
- **Classification**: $y \in \{1, -1\}$, prediction $\hat{y} = \text{sign}(f(x))$
  - loss of the form $\ell(y, f) = \ell(yf)$
  - true loss $\ell(yf) = \delta_{(yf<0)}$
  - useful convex loss





Regularization

- **Main goal**: Avoid over-fitting
- **Two main lines of work**
  - Euclidean and Hilbertian norms (i.e., ℓ_2-norms)
    - Possibility of non linear predictors
    - Non parametric supervised learning and kernel methods
    - Well developed theory and algorithms (see, e.g., Wahba, 1990; Shawe-Taylor and Cristianini, 2004)
  - Sparsity-inducing norms
    - Usually restricted to linear predictors on vectors $f(x) = w^\top x$
    - Main example: ℓ_1-norm $\|w\|_1 = \sum_{i=1}^{p} |w_i|$
    - Perform model selection as well as regularization
    - Theory and algorithms "in the making"

Sparse Linear Estimation with ℓ_1-norm

- The general setting $f : \mathcal{X} \to \mathcal{Y}$

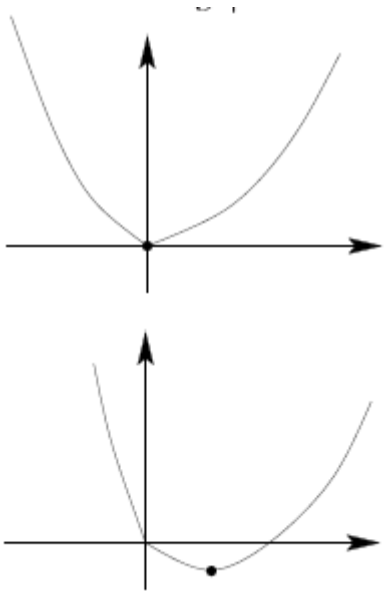
$$\sum_n \ell(f(x_i), y_i) \quad + \quad \Omega(f)$$

- Sparse linear estimation with ℓ_1-norm

$$f(x) = w^\top x$$

$$\Omega(f) = \|w\|_1 = \sum_{i=1}^{p} |w_i|$$

Why ℓ_1-norm leads to sparsity?



- **Example 1**: quadratic problem in 1D

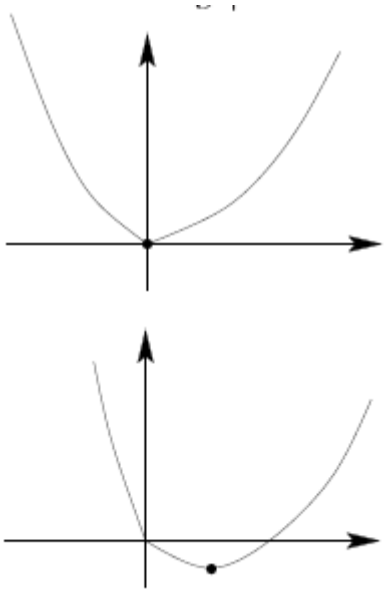
$$\min_x \frac{1}{2}x^2 - xy + \lambda|x|$$

- Piecewise quadratic function with a kink at zero
  - Derivative at 0+: $g_+ = -y + \lambda$
  - Derivative at 0-: $g_- = -y - \lambda$
  - $x = 0$ is the solution iff $g_+ \geq 0,\ g_- \leq 0$ (i.e.: $|y| \leq \lambda$)
  - $x \geq 0$ is the solution iff $g_+ \leq 0$ (i.e.: $y \geq \lambda$) $x^* = y - \lambda$
  - $x \leq 0$ is the solution iff $g_- \geq 0$ (i.e.: $y \leq -\lambda$) $x^* = y + \lambda$
- Solution is: $x^* = \text{sign}(y)(|y| - \lambda)_+$ Soft Thresholding

Why ℓ_1-norm leads to sparsity?

- **Example 1**: quadratic problem in 1D

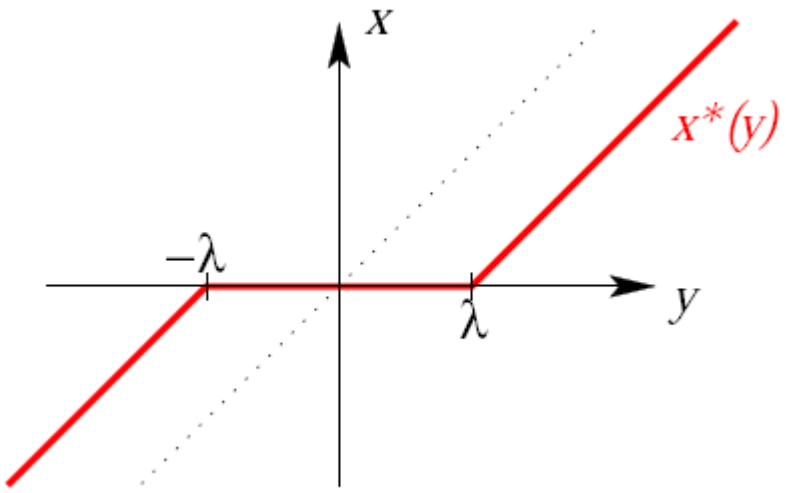
$$\min_x \frac{1}{2}x^2 - xy + \lambda|x|$$



- Piecewise quadratic function with a kink at zero
- Solution is:

Soft Thresholding

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

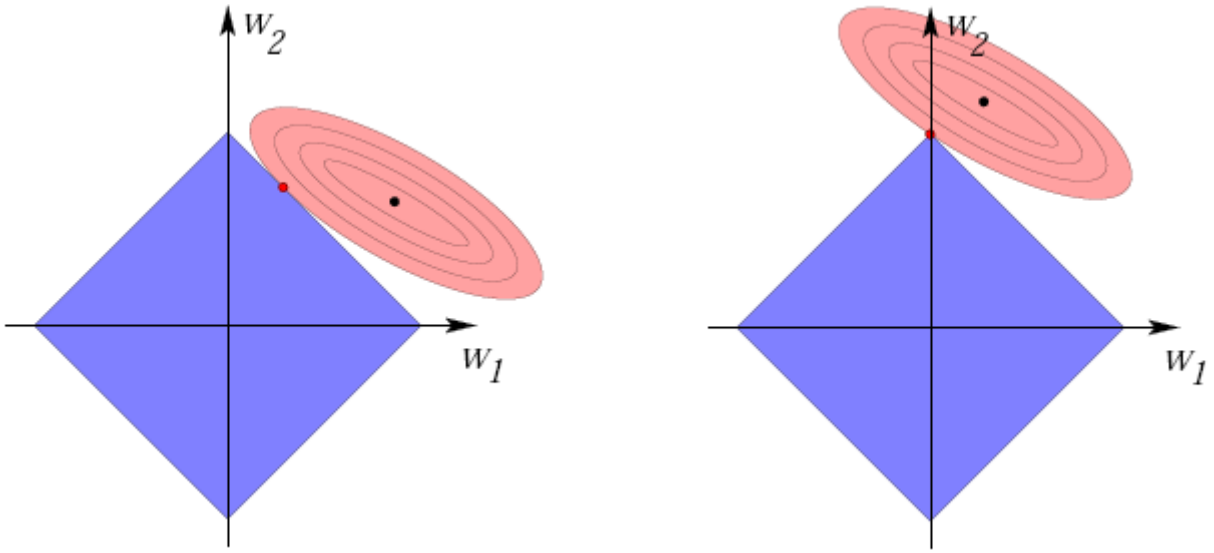


Why ℓ_1-norm leads to sparsity?

- **Example 2**: minimize quadratic function $Q(w)$ subject to

$$\|w\|_1 \leq T$$

- Geometric Interpretation
  - Penalizing is "equivalent" to constraining



Sparse Linear Estimation with ℓ_1-norm

- **Data**: covariates $x_i \in \mathbb{R}^p$, response $y_i \in \mathcal{Y}$, $i = 1, 2, \ldots n$
- Minimize over loadings/weights $w \in \mathbb{R}^p$

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

- Square loss
  - Basis pursuit in signal processing (Chen et al., 2001)
  - Lasso in statistics/machine learning (Tibshirani, 1996)

LASSO

- a loop of rope that is designed to be thrown around a target and tighten when pulled. It is a well-known tool of the American cowboy.

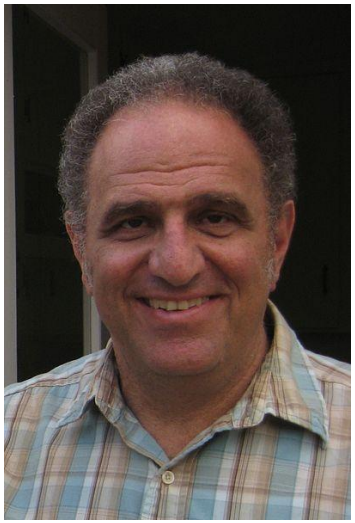


Revisit the Diabetes Study

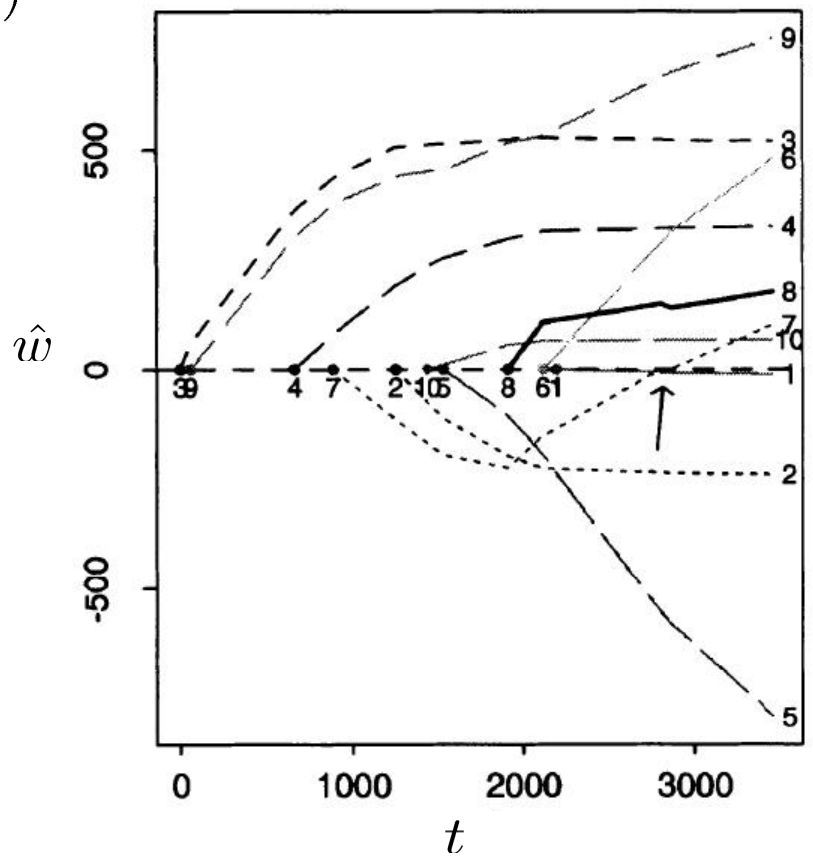
- **Lasso**: least absolute shrinkage and selection operator

$$\min_{w} L(y, w^\top X) = \sum_i (y_i - w^\top x_i)^2$$

$$\text{s.t.: } \|w\|_1 \le t$$



Robert Tibshirani
Dept. Stat., Stanford U.



Nonsmooth convex analysis & optimization

- Analysis
  - optimal conditions
- Optimization
  - algorithms

Optimal conditions for smooth optimization – Zero Gradient

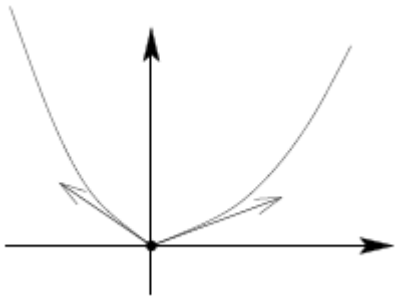
- Example:
$$\min_w J(w) = \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \frac{\lambda}{2}\|w\|_2^2$$
  - gradient
$$\nabla J(w) = \sum_i \nabla_w \ell(y_i, w^\top x_i) x_i + \lambda w$$
  - If squared loss
$$\sum_i \ell(y_i, w^\top x_i) = \frac{1}{2}\|y - Xw\|_2^2$$
    - gradient
$$\nabla J(w) = -X^\top (y - Xw) + \lambda w$$
    - solution
$$w = (\lambda I + X^\top X)^{-1} X^\top y$$
- But ℓ_1-norm is non-differentiable
  - Can't compute the gradient => subgradient (directional derivatives)

Directional Derivatives

- Directional derivative in direction Δ at w :

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

- Always exist when J is convex and continuous
- **Main idea**: in non-smooth settings, may need to look at all directions



- **Proposition**: J is differentiable at w iff $\Delta \mapsto \nabla J(w, \Delta)$ is linear

$$\nabla J(w, \Delta) = \nabla J(w)^\top \Delta$$

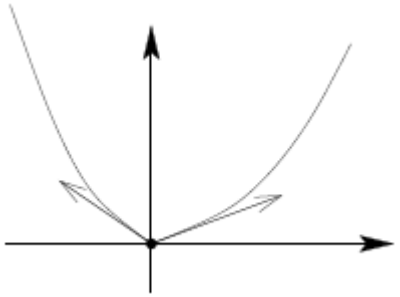
Optimal conditions for convex functions

- Unconstrained minimization
  - **Proposition**: w is optimal iff

$$\forall \Delta \in \mathbb{R}^p : \ \nabla J(w, \Delta) \geq 0$$

  - i.e., function value goes up in all directions

- Reduces to zero-gradient for smooth problems



Directional derivative for ℓ_1-norm

- Function

$$J(w) = \sum_i \ell(y_i, w^\top x_i) + \lambda \|w\|_1 = L(y, Xw) + \lambda \|w\|_1$$

- ℓ_1-norm:

$$|w + \epsilon\Delta\|_1 - \|w\|_1 = \sum_{j, w_j \neq 0} (|w_j + \epsilon\Delta_j| - |w_j|) + \sum_{j, w_j = 0} |\epsilon\Delta_j|$$

- Thus (separability of optimal conditions)

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

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

Directional derivative for ℓ_1-norm

- **General loss**: w is optimal if and only if for all $j = 1, 2, \ldots, p$

$$w_j \neq 0 \quad \Rightarrow \nabla L(w)_j + \lambda \text{sign}(w_j) = 0$$

$$w_j = 0 \quad \Rightarrow |\nabla L(w)_j| \leq \lambda$$

- **Squared loss**: $L(y, Xw) = \sum_i (y_i - w^\top x_i)^2$

$$\nabla L(w)_j = -X_j^\top (y - Xw)$$

- X_j is the j-th column of X

First-order methods for convex optimization – smooth optimization

- **Gradient descent**:

$$w_{t+1} = w_t - \alpha_t \nabla J(w_t)$$

  - with line search: search for a descent α_t
  - with fixed step size, e.g., $\alpha_t = a(t+b)^{-1}$

- Convergence of $f(w_t)$ to $f^*(w) = \min_w f(w)$
  - Depends on the condition number of the optimization number (i.e., correlation within variables)
- **Coordinate descent**:
  - Similar properties

Regularized problems – proximal methods

- Gradient descent as a proximal method

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

$$= w_t - \frac{1}{\mu}\nabla L(w_t)$$

- Regularized problems of the form $\min_w L(w) + \lambda\Omega(w)$

$$w_{t+1} = \arg\max_w L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda\Omega(w_t) + \frac{\mu}{2}\|w - w_t\|_2^2$$

$$= \text{SoftThreshold}(w_t - \frac{1}{\mu}\nabla L(w_t))$$

- Similar convergence rates as sooth optimization
  - Acceleration methods (Nestrov, 2007; Beck & Teboulle, 2009)

More on Proximal Mapping

- The **proximal mapping** (or proximal operator) of a convex function h is

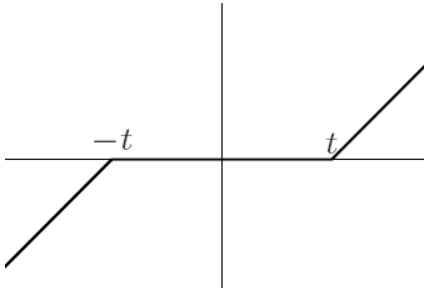
$$\text{prox}_h(x) = \arg\min_{\mu}(h(\mu) + \frac{1}{2}\|\mu - x\|_2^2)$$

- Examples:
  - $h(x) = 0$: $\text{prox}_h(x) = x$
  - $h(x) = I_C(x)$ (indicator function of C): a projection on C

$$\text{prox}_h(x) = P_C(x) = \arg\min_{\mu \in C} \|\mu - x\|_2^2$$

  - $h(x) = t\|x\|_1$: a shrinkage (soft-threshold) operation

$$\text{prox}_h(x)_i = \begin{cases} x_i - t & x_i \geq t \\ 0 & |x_i| \leq t \\ x_i + t & x_i \leq -t \end{cases}$$



More on Proximal Gradient Methods

- **Unconstrained problem** with cost function split in two parts

$$\min_x f(x) = g(x) + h(x)$$

  - g is convex, differentiable
  - h closed, convex, possibly nondifferentiable; prox_h is inexpensive

- **Proximal gradient algorithm**:

$$x^{(k+1)} = \text{prox}_{t_k h}\left(x^{(k)} - t_k \nabla g(x^{(k)})\right)$$

  - t_k is step size, constant or determined by line search

More on Proximal Gradient Methods

$$x^{(k+1)} = \text{prox}_{t_k h}\left(x^{(k)} - t_k \nabla g(x^{(k)})\right)$$

- From definition of proximal operator

$$x^{(k+1)} = \arg\min_{\mu}\left(h(\mu) + \frac{1}{2t_k}\|\mu - x^{(k)} + t_k \nabla g(x^{(k)})\|_2^2\right)$$

$$= \arg\min_{\mu}\left(h(\mu) + g(x^{(k)}) + \nabla g(x^{(k)})^\top(\mu - x^{(k)}) + \frac{1}{2t_k}\|\mu - x^{(k)}\|_2^2\right)$$

  - i.e., minimizes $h(\mu)$ plus a simple quadratic local model of $g(\mu)$ around $x^{(k)}$

Examples

$$\min_x g(x) + h(x)$$

- **Gradient method**: $h(x) = 0$

$$x^{(k+1)} = x^{(k)} - t_k \nabla g(x^{(k)})$$

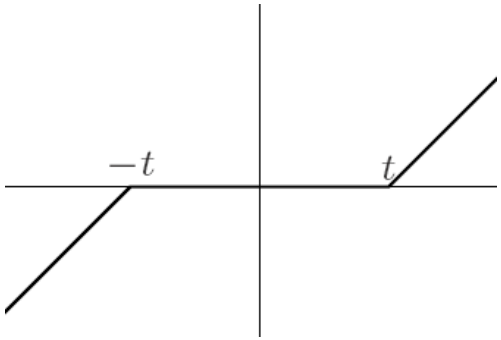
- **Gradient projection method**: $h(x) = I_C(x)$

$$x^{(k+1)} = P_C\left(x^{(k)} - t_k \nabla g(x^{(k)})\right)$$

- **Iterative soft-thresholding**: $h(x) = t\|x\|_1$

$$x^{(k+1)} = \text{prox}_{t_k h}\left(x^{(k)} - t_k \nabla g(x^{(k)})\right)$$

$$\text{prox}_h(x)_i = \begin{cases} x_i - t & x_i \geq t \\ 0 & |x_i| \leq t \\ x_i + t & x_i \leq -t \end{cases}$$



η -Trick for ℓ_1-norm

- Variational form of the ℓ_1 -norm

$$\|w\|_1 = \min_{\eta \geq 0} \frac{1}{2} \sum_{i=1}^{p} \left(\frac{w_i^2}{\eta_i} + \eta_i \right)$$

- **Alternating minimization**
  - For η , closed-form solution $\eta_i = |w_i|$
  - For w, weighted squared ℓ_2 -norm regularized problem
  - **Caveat**: lack of continuity around $(w_i, \eta_i) = (0, 0)$

QP Formulation

- For the special case with square loss

$$\min_{w} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$$

  - is equivalent to ($w = w^+ - w^-$)

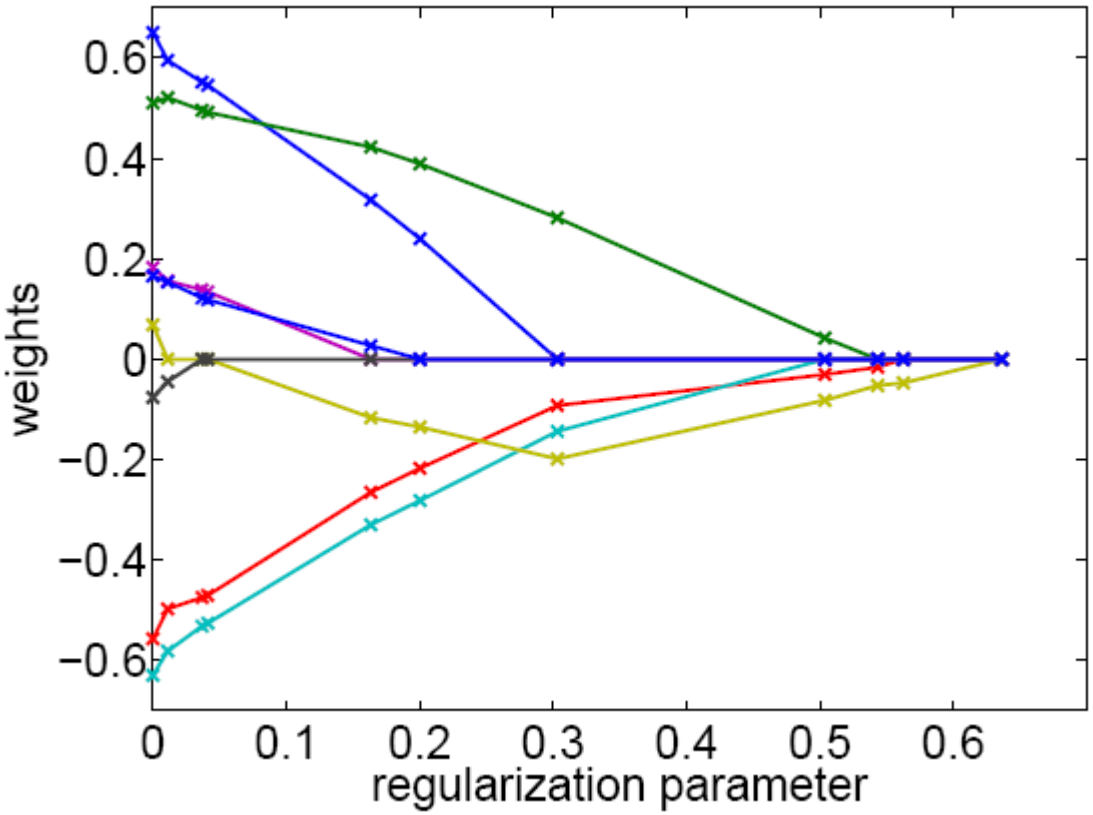
$$\min_{w^+, w^-} \frac{1}{2}\|y - X(w^+ - w^-)\|_2^2 + \lambda \sum_{j=1}^{p}(w_j^+ + w_j^-)$$

$$\text{s.t.: } w^+ \geq 0,\ w^- \geq 0$$

  - generic toolboxes apply, but normally very slow!

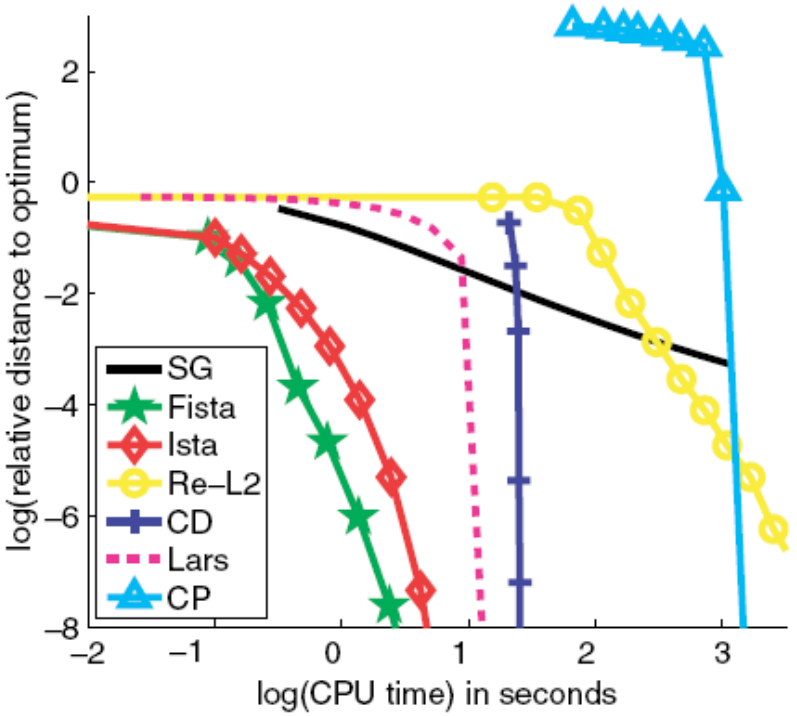
Piecewise linear solution paths

- **Lasso:**

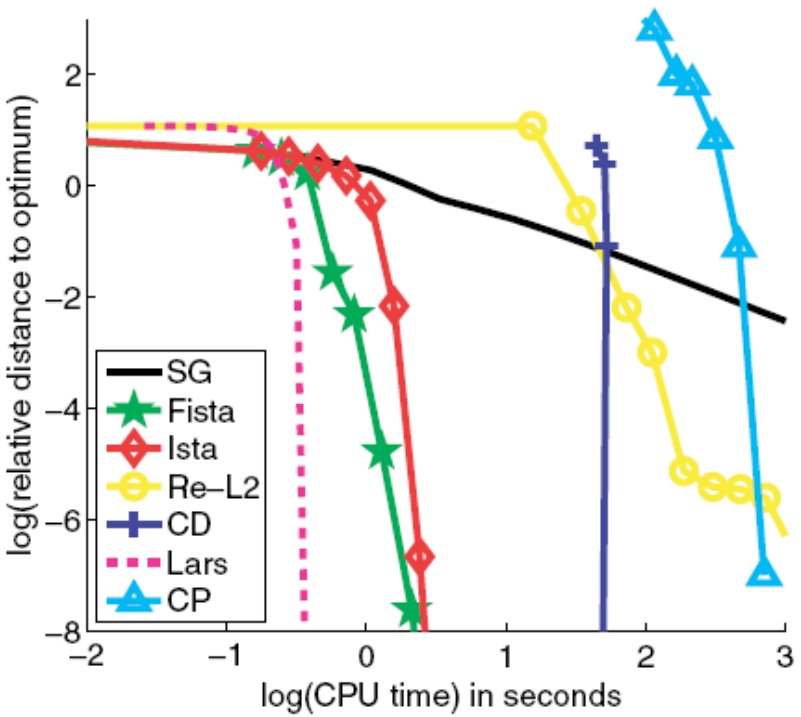


Comparison on Algorithms for Lasso

$$n = 2000, p = 10,000$$



(a) corr: low, reg: low



(b) corr: low, reg: high

- SG: sub-gradient descent
- Ista: simple proximal methods
- Fista: accelerated version of Ista
- Re-L2: reweighted-least square
- CP: cone programming
- QP: quadratic programming
- Lars: least angle regression
- CD: coordinate descent

Alternative sparse methods – Greedy methods

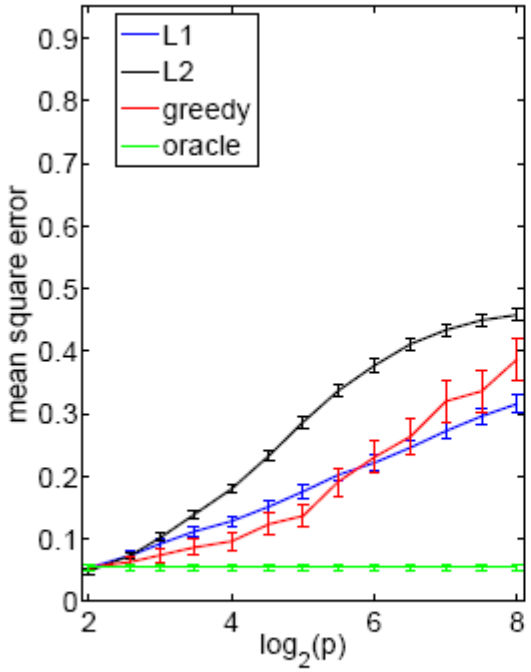
- Forward selection
- Forward-backward selection

- Non-convex method
  - Harder to analyze
  - Simpler to implement
  - Problems of stability

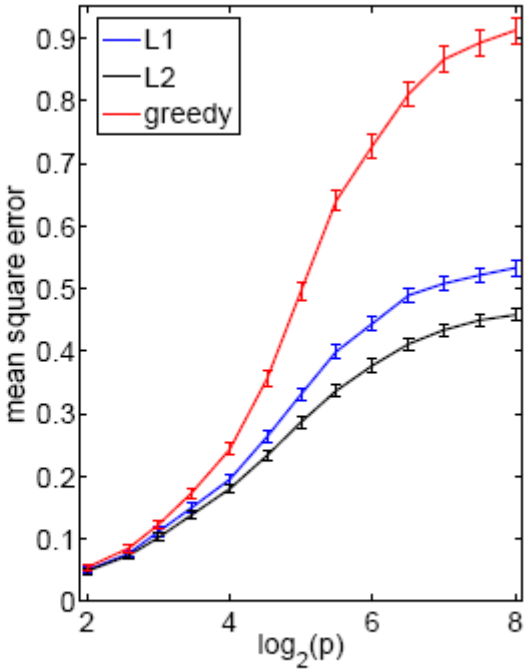
- Positive theoretical results (Zhang, 2009, 2008a)
  - Similar sufficient conditions than for the Lasso

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)

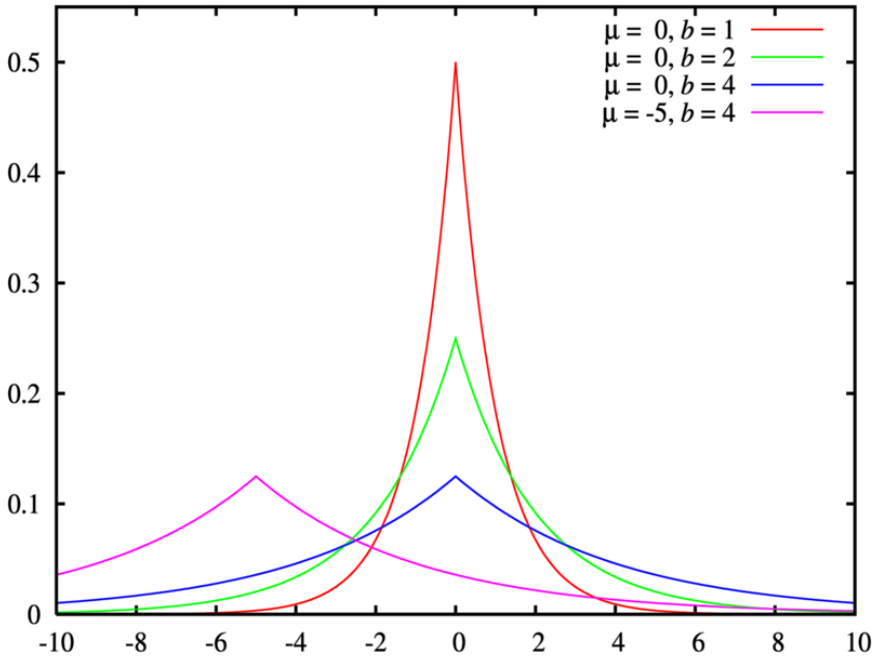
Alternative sparse methods – Bayesian methods

- **Lasso**:

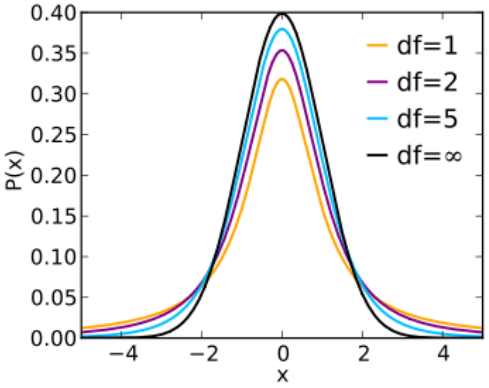
$$\min_w \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_1$$

  - Equivalent to MAP estimate of a Gaussian likelihood model with Laplace prior
  - However, posterior puts zero weights on exact zeros

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(- \frac{|x - \mu|}{b} \right)$$



Alternative sparse methods – Bayesian methods

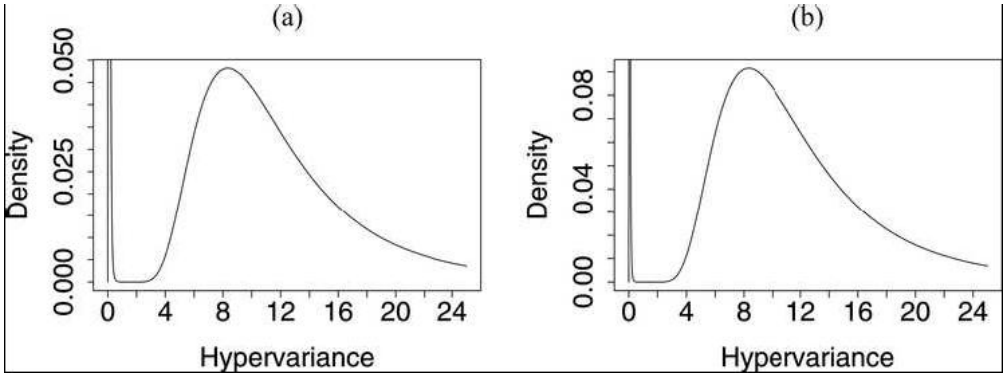


- **Heavy-tailed priors as a proxy to sparsity**
  - Student distributions (Caron and Doucet, 2008)
  - Generalized hyperbolic priors (Archambeau and Bach, 2008)
  - Instance of automatic relevance determination (Neal, 1996)

- **Spike and Slab** (Ishwaran and Rao, 2005)
  - Mixtures of "Diracs" and another absolutely continuous distributions

$$p(\sigma^2) = \alpha\delta(\sigma^2 = v_0) + (1-\alpha)\text{Gamma}(a,b)$$

where v_0 is a small near zero value



- **Less theory than frequentist methods**

Relevance Vector Machine

- An additive noise model

$$t_n = y(x_n; w) + \epsilon_n, \quad \text{where } \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

  - for generalized linear model, we have

$$y(x; w) = \sum_{i=1}^{M} w_i \phi_i(x) = w^\top \phi(x)$$

- Thus, the likelihood model is

$$p(t_n|x_n) = \mathcal{N}(y(x_n), \sigma^2) = \mathcal{N}(w^\top \phi(x_n), \sigma^2)$$

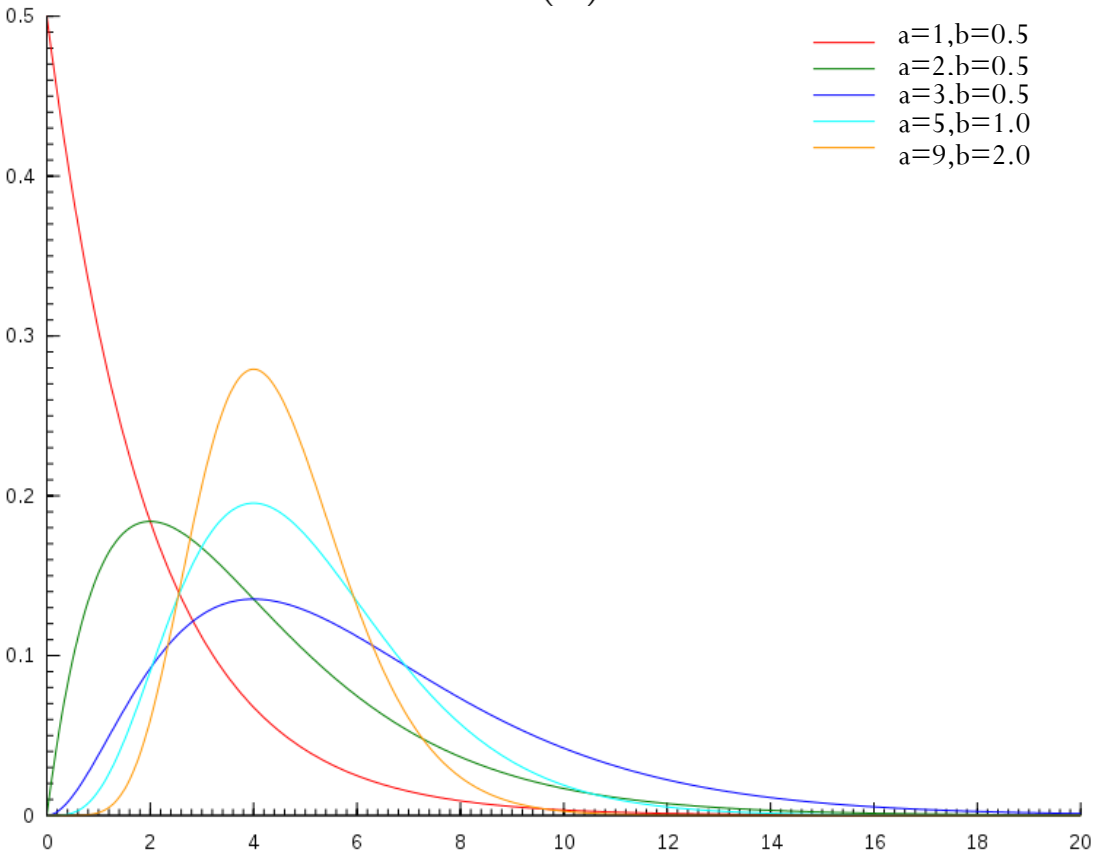
- To be Bayesian, assume common zero-mean Gaussian prior

$$p(w|\alpha) = \prod_{i=0}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1})$$

$$p(\alpha) = \prod_{i=0}^{M} \text{Gamma}(\alpha_i|a, b) \qquad p(\beta) = \text{Gamma}(\beta|c, d) \text{ with } \beta = \sigma^{-2}$$

Gamma Distribution

$$p(x;a,b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-ba}$$



Inference

- Posterior distribution

$$p(w, \alpha, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|w, \alpha, \sigma^2)p(w, \alpha, \sigma^2)}{p(\mathbf{t})}$$

- Prediction

$$p(t_*|\mathbf{t}) = \int p(t_*|w, \alpha, \sigma^2)p(w, \alpha, \sigma^2|\mathbf{t})dwd\alpha d\sigma^2$$

- The intractable posterior can be approximated as

$$p(w, \alpha, \sigma^2|\mathbf{t}) \approx p(w|\alpha, \sigma^2, \mathbf{t})\delta(\alpha_{\text{MAP}}, \sigma^2_{\text{MAP}})$$

where $p(w|\alpha, \sigma^2, \mathbf{t}) = \frac{p(\mathbf{t}|w, \sigma^2)p(w|\alpha)}{p(\mathbf{t}|\alpha, \sigma^2)} = \mathcal{N}(\mu, \Sigma)$ is Normal; and

$$(\alpha_{\text{MAP}}, \sigma^2_{\text{MAP}}) = \arg\max_{(\alpha,\sigma^2)} p(\alpha, \sigma^2|\mathbf{t})$$

Relevance Vector Machine

- Sparse MAP estimates

$$(\alpha_{\text{MAP}}, \sigma^2_{\text{MAP}}) = \arg \max_{(\alpha,\sigma^2)} \log p(\alpha, \sigma^2 | \mathbf{t})$$

- For the non-informative prior, we have

$$(\alpha_{\text{MAP}}, \sigma^2_{\text{MAP}}) = \arg \max_{(\alpha,\sigma^2)} \log p(\mathbf{t} | \alpha, \sigma^2)$$

  - a QP problem due to the conjugacy of normal distributions
  - The update equations are

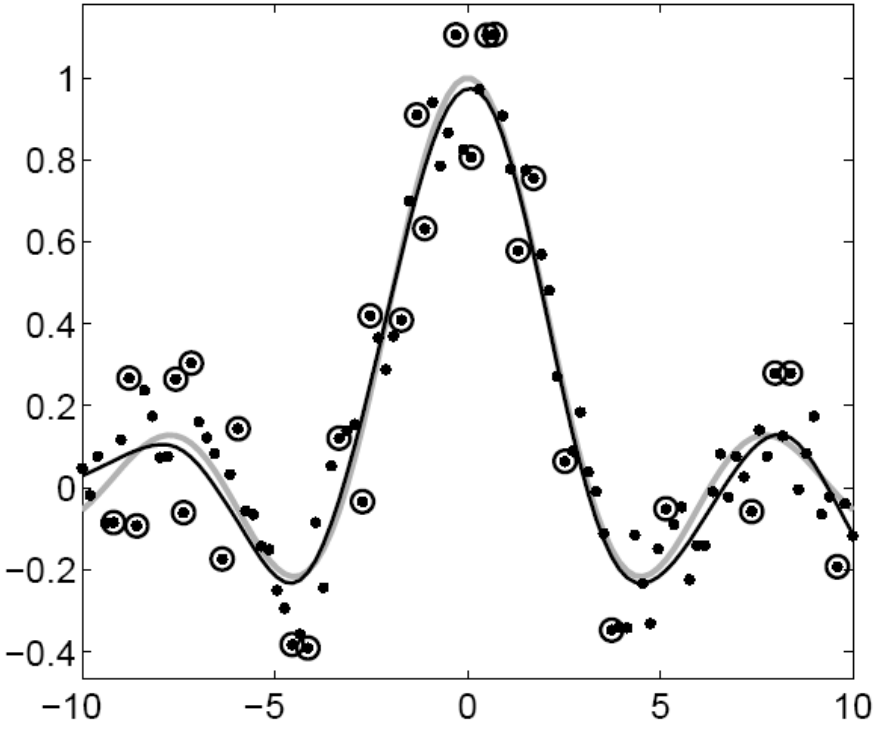
$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}, \text{ where } \gamma_i = 1 - \alpha_i \Sigma_{ii} \in (0, 1)$$

$$(\sigma^2)^{new} = \frac{\|\mathbf{t} - \Phi\mu\|_2^2}{N - \sum_i \gamma_i}$$

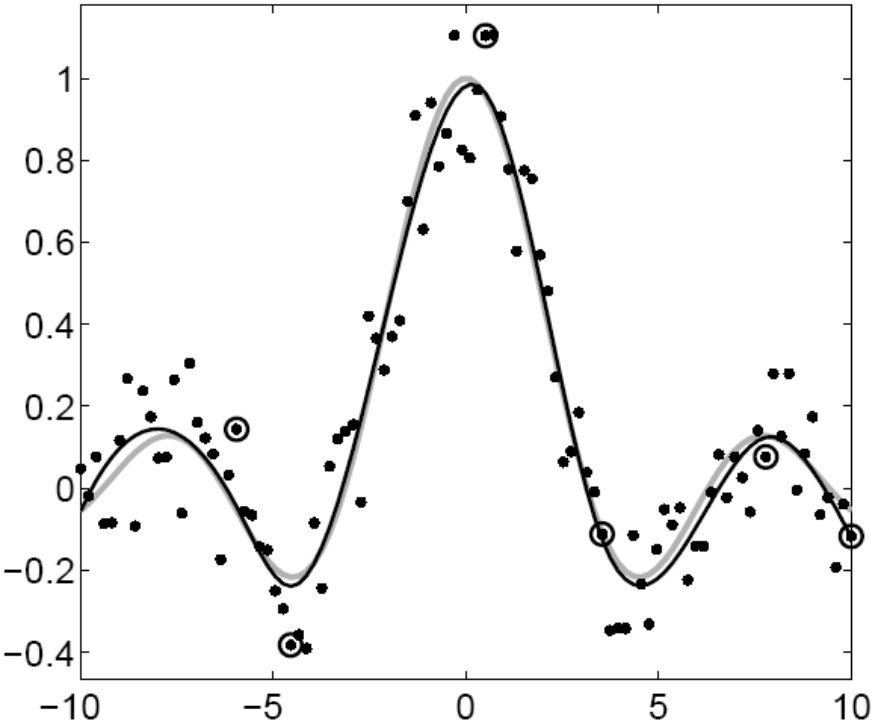
$$\Sigma = (\sigma^{-2}\Phi^\top\Phi + \text{diag}(\alpha))^{-1}$$

Example – Regression

- Simulation form function $f(x) = \dfrac{\sin(x)}{x}$ with uniform noise at [-0.2, 0.2]



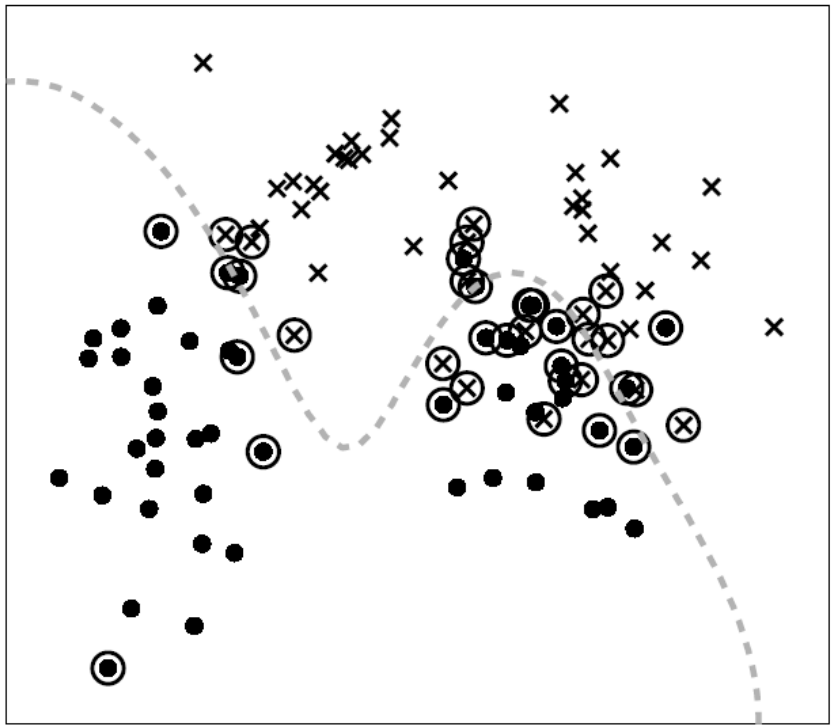
support vector regression



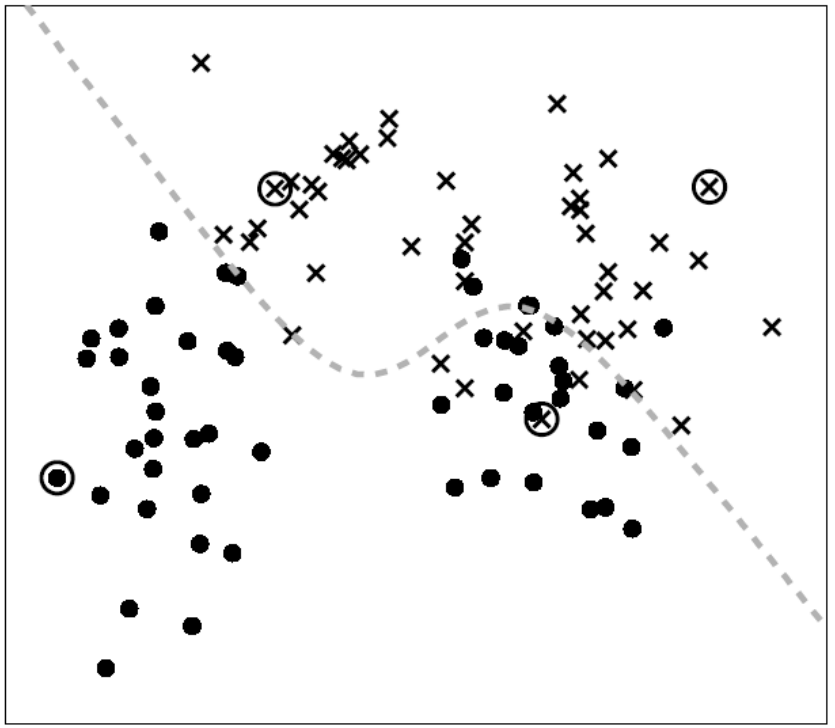
relevance vector regression

Example – Classification

- Synthetic data from mixture of two Gaussians
- Gaussian RBF kernel



support vector machines



relevance vector classification

Summary -- ℓ_1-norm Regularization

- Leads to non-smooth optimization
  - analysis through directional derivatives or subgradients
  - optimization may or may not take advantage of sparsity
- Allows high-dimensional inference
- Interesting problems:
  - Stable variable selection
  - Weaker sufficient conditions (for weaker results)
  - Estimation of regularization parameter (all bounds depend on the unknown noise variance $\sigma 2$)

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 Buhlmann, 2006; Banerjee et al., 2008)
  - Sparsity on matrices
- Sparse methods are not limited to variable selection in a linear model
  - Kernel learning (Bach et al., 2008)

Regularization with Groups of Variables

- Assume $\{1, 2, \ldots p\}$ is partitioned into m groups $G_1, G_2, \ldots G_m$
- Regularization:

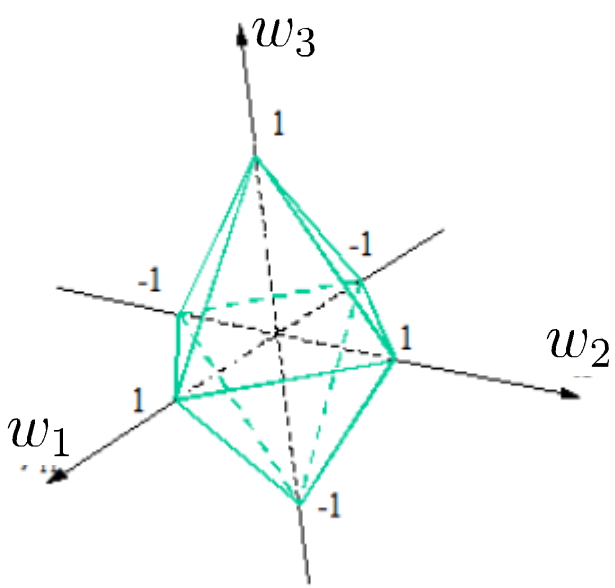
$$\Omega(w) = \sum_{i=1}^{m} \|w_{G_i}\|_2$$

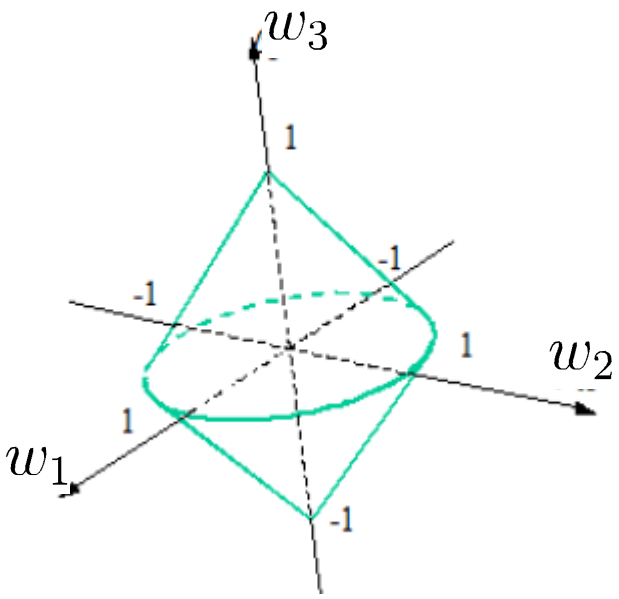
- Induces groupwise sparsity
  - Some groups entirely set to zero
  - No zeros within group

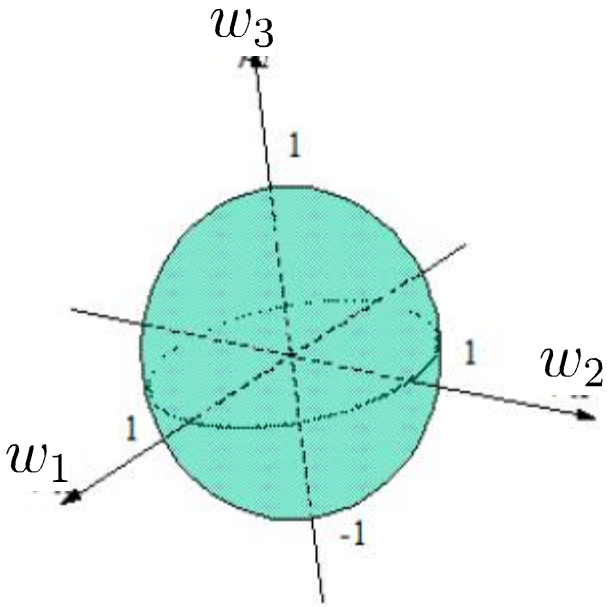
Regularization with Groups of Variables

$$\Omega(w) = \sum_{i=1}^{m} \|w_{G_i}\|_2$$

- E.g.: $\|(w_1, w_2)\|_2 + \|w_3\|_2 \leq 1$







Group Lasso

- **Opt. problem**:

$$\min_w \sum_i (y_i - w^\top x_i)^2 + \lambda \sum_{i=1}^m \sqrt{p_i} \|w_{G_i}\|_2$$

- Optimal condition:
  - **Proposition**: w is optimal iff $\forall j = 1, 2, \ldots, m$

$$w_{G_j} \neq 0 \quad \Rightarrow -X_{G_j}^\top (y - Xw) + \frac{\lambda \sqrt{p_j} w_{G_j}}{\|w_{G_j}\|_2} = 0$$

$$w_{G_j} = 0 \quad \Rightarrow \|X_{G_j}^\top (y - Xw)\|_2 \leq \lambda \sqrt{p_j}$$

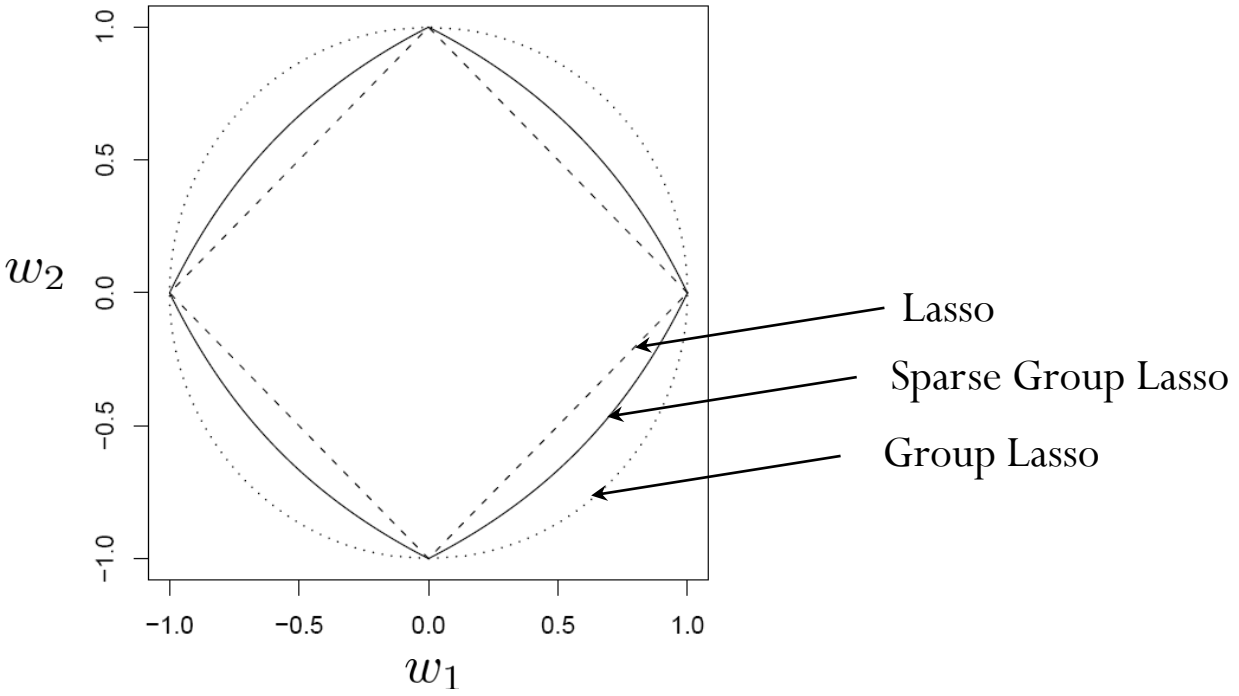
  - p_j is the number of features in group j.
- Coordinate descent algorithm can be used to solve it.

Sparse Group Lasso

- **Opt. Problem**:

$$\min_{w} \sum_{i} (y_i - w^\top x_i)^2 + \lambda \sum_{i=1}^{m} \|w_{G_i}\|_2 + \gamma \|w\|_1$$

  - the single group case:



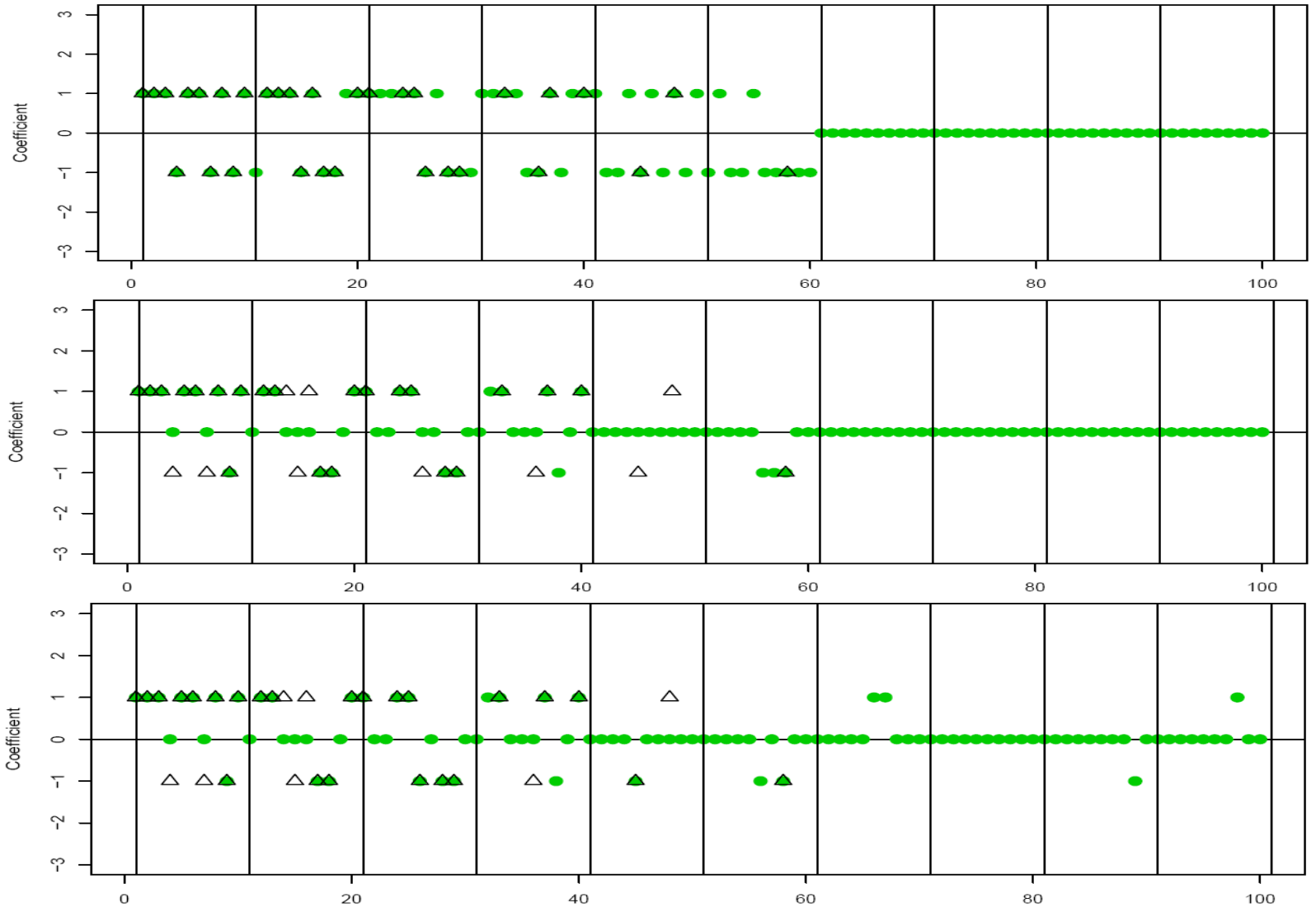
[Friedman et al., 2010]

Example

Group Lasso

Sparse Group Lasso

Lasso



Kernel methods: regularization with ℓ_2 -norm

- Feature mapping $\Phi : \mathcal{X} \to \mathbb{R}^p$
- Model $f(x) = w^\top \Phi(x)$
  - Linear in parameters; nonlinear in data
- Optimization problem:

$$\min_w \sum_i \|y_i - w^\top \Phi(x_i)\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$$

- Representer theorem (Kimeldorf & Wahba, 1971): solution is of the form $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - Equivalent to solve $\min_\alpha \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^\top K \alpha$
  - K is the kernel matrix

$$K_{ij} = \kappa(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$$

Multiple Kernel Learning (MKL)

- Multiple feature maps / kernels on inputs
  - p "feature maps": $\Phi_j : \mathcal{X} \to \mathcal{F}_j, \;\; j = 1, 2, \ldots, p$
  - **The model** (predictor):

$$f(x) = w_1^\top \Phi_1(x) + w_2^\top \Phi_2(x) + \cdots + w_p^\top \Phi_p(x)$$

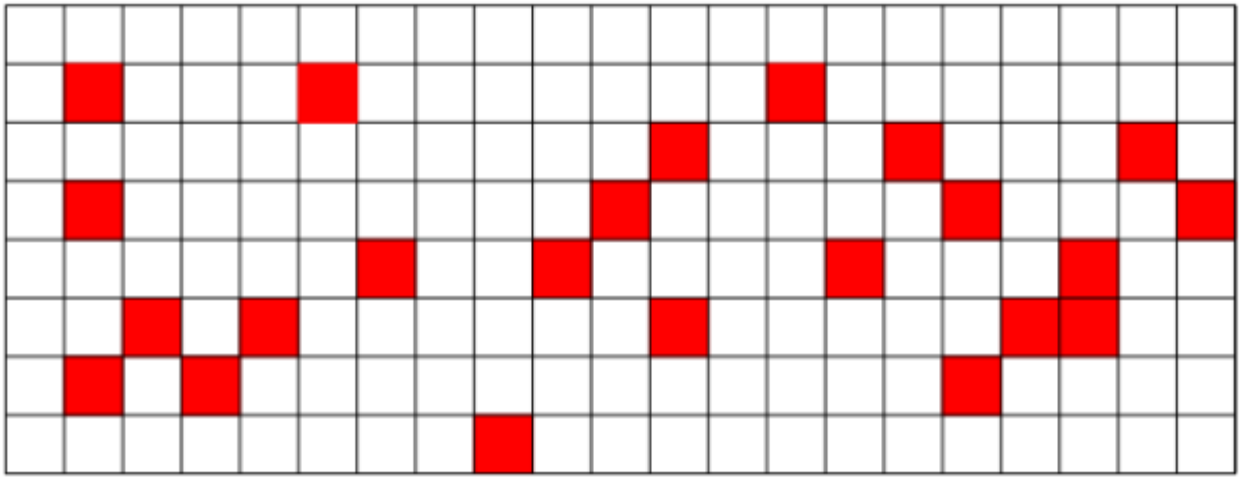
$$x \longrightarrow \begin{matrix} \Phi_1(x)^\top & w_1 \\ \vdots & \vdots \\ \Phi_j(x)^\top & w_j \\ \vdots & \vdots \\ \Phi_p(x)^\top & w_p \end{matrix} \longrightarrow w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x)$$

  - Generalized linear models (Hastie & Tibshirani, 1990)

Sparsity for Matrices

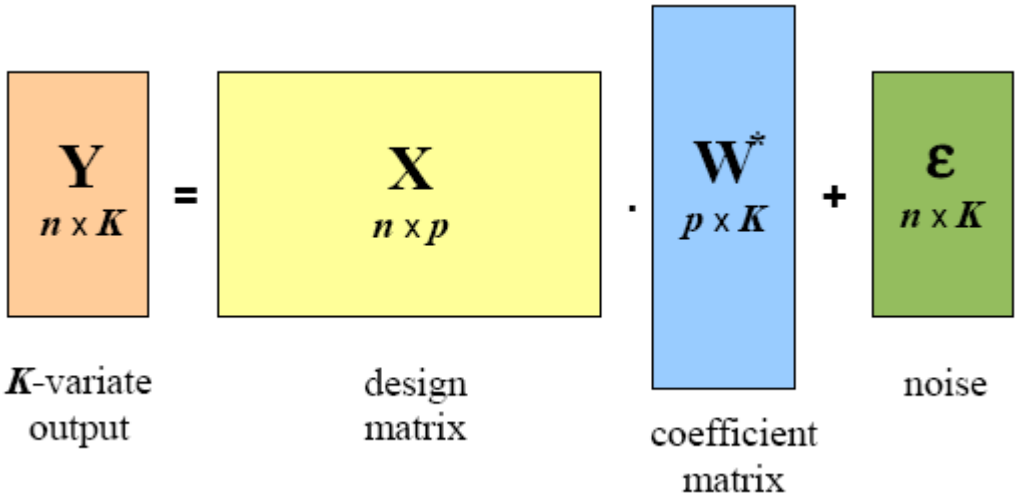
Learning on Matrices – Collaborative Filtering

- 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 Z that describes the known ratings of some customers for some movies
- Goal: complete the matrix.



Learning on Matrices
– Multivariate Regression/Classification

- Multivariate linear regression



- Multiclass linear classification

$$\min_{\mathrm{W}} \sum_{d=1}^{n} \frac{1}{n} \ell(w_1^\top x_d, \ldots, w_M^\top x_d, y_d)$$

  - where $y_d \in \{0,1\}^K$ and ℓ is the loss, e.g., logistic loss

Learning with Matrices – Multi-task Learning

- K prediction tasks on the same covariates $x \in \mathbb{R}^P$
  - Each model parameterized by $w_k \in \mathbb{R}^P$
  - Empirical risks:

$$L_k(w_k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k(w_k^\top x_i^k, y_i^k)$$

  - All the parameters form a matrix

$$W = [w_1, \ldots, w_K] = \begin{bmatrix} w_1^1 & \ldots & w_K^1 \\ \vdots & w_k^j & \vdots \\ w_1^p & \ldots & w_K^p \end{bmatrix} = \begin{bmatrix} w^1 \\ \vdots \\ w^p \end{bmatrix}$$

- Many applications:
  - Multi-category classification (one task per class)

Example – Image Denoising

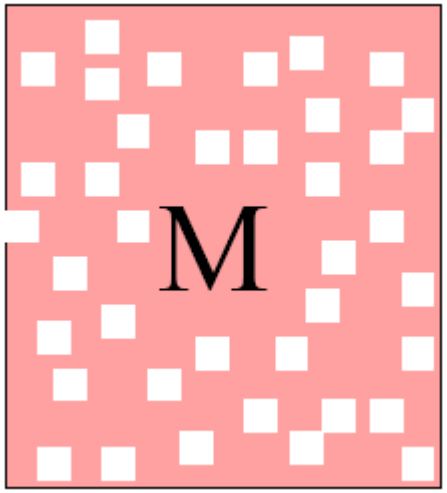
- Simultaneously denoise all patches of a given image
- Example from Mairal et al. (2009)



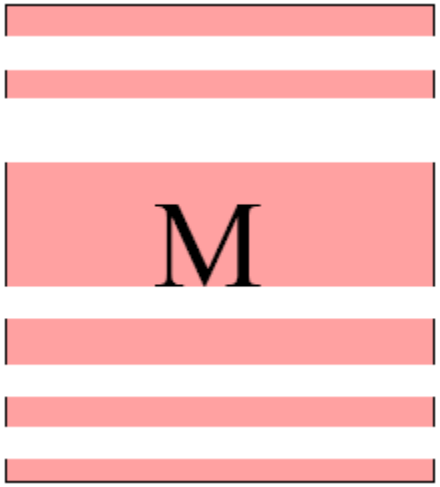


Two types of sparsity of matrices

- **Type 1 of sparsity**:
  - Directly on the elements



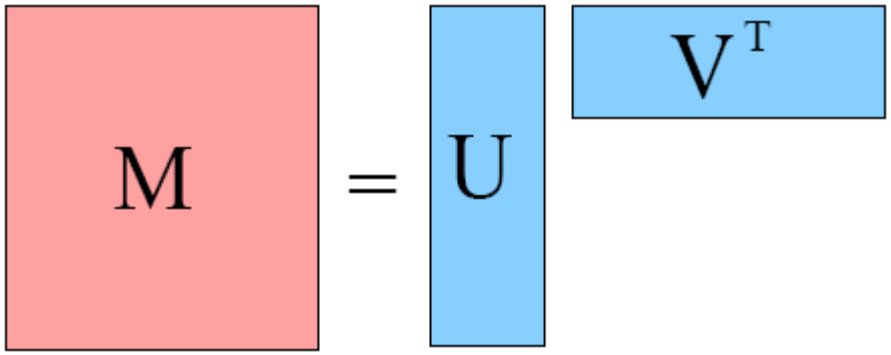
Many elements are zeros



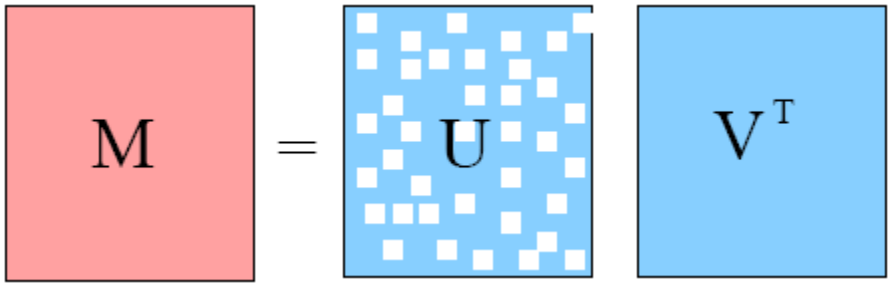
Many rows or columns are zeros

Two types of sparsity of matrices

- **Type 2 of sparsity**:
  - Through a factorization $M = UV^\top$ $U \in \mathbb{R}^{n\times m}$, $V \in \mathbb{R}^{p\times m}$
  - Low-rank sparsity: m is small



  - Sparse decomposition: U sparse



Type 1: Joint Variable Selection in MTL

- Parameters for all the K tasks

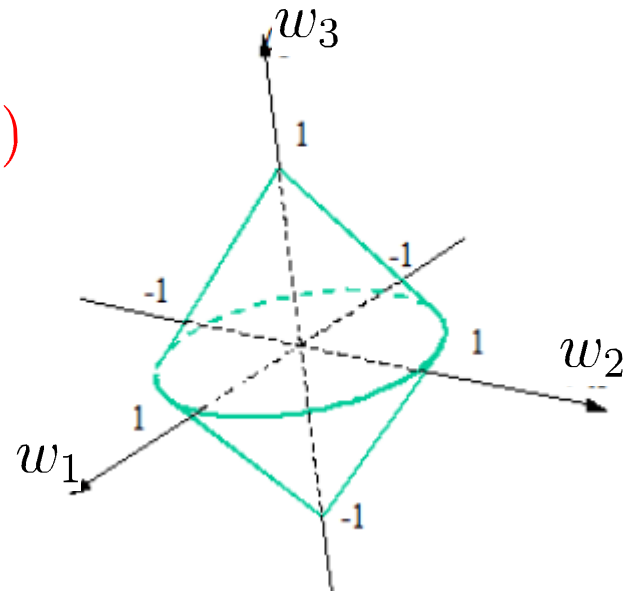
$$W = [w_1, \ldots, w_K] = \begin{bmatrix} w_1^1 & \ldots & w_K^1 \\ \vdots & w_k^j & \vdots \\ w_1^p & \ldots & w_K^p \end{bmatrix} = \begin{bmatrix} w^1 \\ \vdots \\ w^p \end{bmatrix}$$

- Select all variables that are relevant to at least one task

$$\min_W \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k(w_k^\top x_i^k, y_i^k) + \lambda \Omega(W)$$

  - which regularizer?

$$\Omega(W) = \sum_{k=1}^{K} \|w_k\|_2$$



Type 2: Rank constraints and sparsity of the spectrum

- Given a matrix $M \in \mathbb{R}^{n \times p}$
  - Singular value decomposition (SVD): $M = U \text{diag}(\lambda) V^\top$

  where U and V are orthogonal matrices; $\lambda \in \mathbb{R}^m$ are eigenvalues
  - The rank of M is:

$$\text{rank}(M) = \|\lambda\|_0$$

  - Rank of M is the minimum size m of all factorizations $M = UV^\top$ where $U \in \mathbb{R}^{n \times m}$, $V \in \mathbb{R}^{p \times m}$
- Rank constrained learning

$$\min_{W \in \mathbb{R}^{n \times p}} L(W) \quad \text{s.t.:} \quad \text{rank}(W) \leq m$$

Low-rank via Factorization

- **Reduced-rank multivariate regression**

$$\min_{W \in \mathbb{R}^{n \times p}} \|Y - XW\|_F^2 \quad \text{s.t.:} \quad \text{rank}(W) \leq m$$

  - Well studied (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998)
  - Is solved directly using the SVD (by OLS + SVD + projection)

- **General formulation**

$$\min_{U \in \mathbb{R}^{n \times m}, V \in \mathbb{R}^{p \times m}} L(UV^\top)$$

  - Still non-convex but convex w.r.t. U and V separately
  - Optimization by alternating procedures

Trace-norm Relaxation

- With SVD $M = U\mathrm{diag}(\lambda)V^\top \quad \mathrm{rank}(M) = \|\lambda\|_0$
  - 0-norm can be relaxed by 1-norm $\|\lambda\|_1$
  - This is the trace norm, denoted by $\|M\|_{\mathrm{tr}} = \|\lambda\|_1$

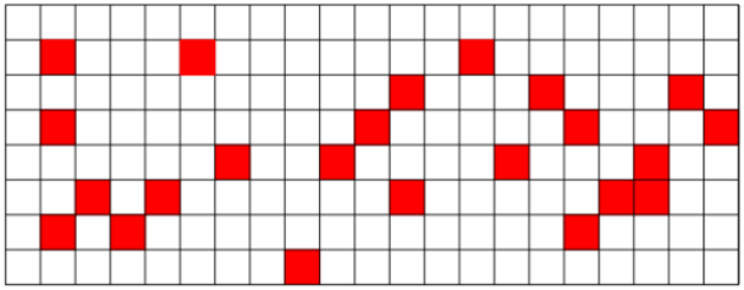
- **Trace-norm regularized learning**

$$\min_{W \in \mathbb{R}^{n \times p}} L(W) + \lambda \|W\|_{\mathrm{tr}}$$

  - Convex problem
  - Can be solved with: proximal methods; Iterative re-weighted Least-squares; etc

Trace-norm and Collaborative Filtering

- **CF as matrix completion** (users as rows; items as columns)

$Y =$ 

- Find a low-rank matrix to reconstruct noisy observations

$$\min_{X \in \mathbb{R}^{n \times p}} \sum_{(i,j) \in S} (X_{ij} - Y_{ij})^2 + \lambda \|X\|_{\text{tr}}$$

  - Semi-definite program (Fazel et al., 2001)
  - Max-margin approaches to CF (Srebro et al., 2005)
  - High-dimensional inference fro noisy matrix completion (Srebro et al., 2005; Candes & Plan, 2009)
  - May recover entire matrix from slightly more entries than the minimum of the two dimensions

References

- Robert Tibshirani (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267-288.
- Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani (2004), Least Angle Regression, Annals of Statistics, 32, 407-451.
- Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). A note on the group lasso and a sparse group lasso. Tech. Report.
- Francis Bach, Rodolphe Jenatton, Julien Mairal and Guillaume Obozinski. Optimization with Sparsity-Inducing Penalties. Foundations and Trends in Machine Learning, Vol. 4, No. 1 (2011)

- Software:
  - SPAMS (SPArse Modeling Software): http://www.di.ens.fr/willow/SPAMS/

References

- A monograph

Foundations and Trends® in
Machine Learning
Vol. 4, No. 1 (2011) 1–106
© 2012 F. Bach, R. Jenatton, J. Mairal
and G. Obozinski
DOI: 10.1561/2200000015

now
the essence of knowledge

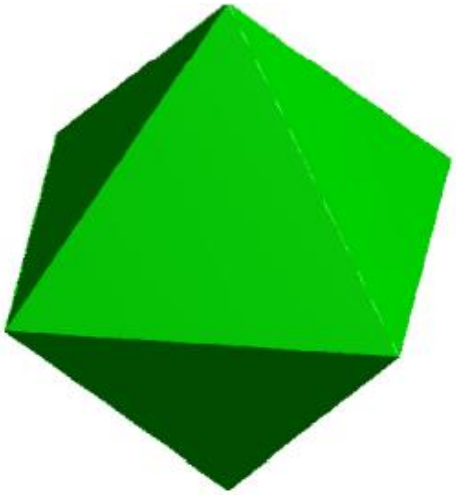
Optimization with Sparsity-Inducing Penalties

By Francis Bach, Rodolphe Jenatton,
Julien Mairal and Guillaume Obozinski

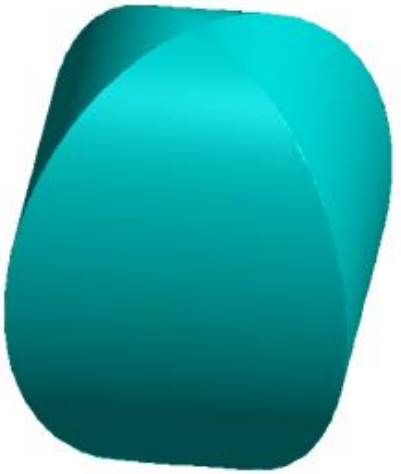
Various Balls in 3D



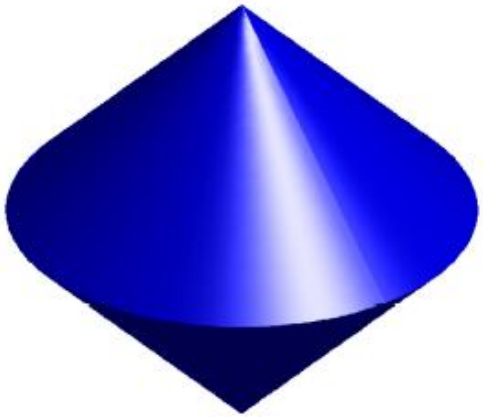
(a) ℓ_2-norm ball.



(b) ℓ_1-norm ball.



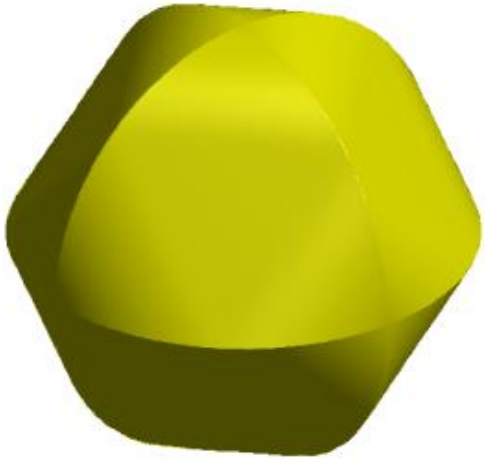
(e) Ω_{union} ball for $\mathcal{G} = \{\{1,3\},\{2,3\}\}$.



(c) ℓ_1/ℓ_2-norm ball: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}_{\{1,2\}}\|_2 + |\boldsymbol{w}_3|$.



(d) ℓ_1/ℓ_2-norm ball: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_2 + |\boldsymbol{w}_1| + |\boldsymbol{w}_2|$.



(f) Ω_{union} ball for $\mathcal{G} = \{\{1,3\},\{2,3\},\{1,2\}\}$.