

# Regularization Algorithms for Learning

Alessandro Verri

DISI, UNIGE

Texas, 10/19/07

- motivation
- setting
- elastic net regularization
  - iterative thresholding algorithms
  - error estimates and parameter choice
- applications

- starting point of many learning schemes is a fixed data representation
- if prediction is the goal, black box solutions are acceptable
- in many problems the primary goal is the identification of the variables/measures relevant for prediction
- it is often the case that variables are dependent

**model:**  $X \times Y$  is endowed with a probability distribution

$$p(y, x) = p(y|x)p(x)$$

**input space:**  $X \subset \mathbb{R}^d$

**regression:**  $Y \in [-M, M]$

**classification:**  $Y \in \{-1, 1\}$

the distribution  $p$  is fixed but unknown

**DATA:** we are given a set of examples, i.e.  $(x_1, y_1), \dots, (x_n, y_n)$   
sampled i.i.d. according to  $p(y, x)$

**model:**  $X \times Y$  is endowed with a probability distribution

$$p(y, x) = p(y|x)p(x)$$

**input space:**  $X \subset \mathbb{R}^d$

**regression:**  $Y \in [-M, M]$

**classification:**  $Y \in \{-1, 1\}$

the distribution  $p$  is fixed but unknown

**DATA:** we are given a set of examples, i.e.  $(x_1, y_1), \dots, (x_n, y_n)$   
sampled i.i.d. according to  $p(y, x)$

- regression function:  $f^*(x) = \mathbb{E}[y|x]$  minimizes

$$\mathcal{E}(f) = \mathbb{E} \left[ |y - f(x)|^2 \right]$$

- Bayes rule:  $f_b(x) = \text{sign}(f^*(x))$  minimizes

$$\mathcal{R}(f) = P(yf(x) \leq 0)$$

- for any  $f$

$$\mathcal{R}(f) - \mathcal{R}(f_b) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f^*)}$$

- regression function:  $f^*(x) = \mathbb{E}[y|x]$  minimizes

$$\mathcal{E}(f) = \mathbb{E} \left[ |y - f(x)|^2 \right]$$

- Bayes rule:  $f_b(x) = \text{sign}(f^*(x))$  minimizes

$$\mathcal{R}(f) = P(yf(x) \leq 0)$$

- for any  $f$

$$\mathcal{R}(f) - \mathcal{R}(f_b) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f^*)}$$

# Hypotheses Space and Dictionary

The search for a solution is restricted to some space of hypotheses  $\mathcal{H}$

- dictionary:  $\mathcal{D} = \{\varphi_\gamma : \mathcal{X} \rightarrow \mathbb{R} \mid \gamma \in \Gamma\}$
- atoms or features:  $\varphi_\gamma$
- hypotheses space:  $\mathcal{H} = \{f \mid f = f_\beta = \sum_{\gamma \in \Gamma} \varphi_\gamma \beta_\gamma\}$

- The atoms are not linearly independent,
- the dictionary can be infinite dimensional,
- the atoms can be seen as measures (features) on the input objects in  $\mathcal{X}$ ,
- the solution is a weighted combination of the features  
$$f_\beta(x) = \sum_{\gamma \in \Gamma} \varphi_\gamma(x) \beta_\gamma$$

# Hypotheses Space and Dictionary

The search for a solution is restricted to some space of hypotheses  $\mathcal{H}$

- dictionary:  $\mathcal{D} = \{\varphi_\gamma : \mathcal{X} \rightarrow \mathbb{R} \mid \gamma \in \Gamma\}$
- atoms or features:  $\varphi_\gamma$
- hypotheses space:  $\mathcal{H} = \{f \mid f = f_\beta = \sum_{\gamma \in \Gamma} \varphi_\gamma \beta_\gamma\}$

- The atoms are not linearly independent,
- the dictionary can be infinite dimensional,
- the atoms can be seen as measures (features) on the input objects in  $\mathcal{X}$ ,
- the solution is a weighted combination of the features  
$$f_\beta(x) = \sum_{\gamma \in \Gamma} \varphi_\gamma(x) \beta_\gamma$$

# Hypotheses Space and Dictionary

The search for a solution is restricted to some space of hypotheses  $\mathcal{H}$

- dictionary:  $\mathcal{D} = \{\varphi_\gamma : \mathcal{X} \rightarrow \mathbb{R} \mid \gamma \in \Gamma\}$
- atoms or features:  $\varphi_\gamma$
- hypotheses space:  $\mathcal{H} = \{f \mid f = \mathbf{f}_\beta = \sum_{\gamma \in \Gamma} \varphi_\gamma \beta_\gamma\}$

- The atoms are not linearly independent,
- the dictionary can be infinite dimensional,
- the atoms can be seen as measures (features) on the input objects in  $\mathcal{X}$ ,
- the solution is a weighted combination of the features

$$\mathbf{f}_\beta(\mathbf{x}) = \sum_{\gamma \in \Gamma} \varphi_\gamma(\mathbf{x}) \beta_\gamma$$

## learning task

given  $(x_1, y_1), \dots, (x_n, y_n)$  find an estimator  $f_n$  such that

$$f_n(x) \sim f^*(x)$$

## An Important Distinction

- The problem of **prediction/generalization** is that of estimating  $f^*$ .
- The problem of **selection** is that of detecting a meaningful  $\beta^*$  (with  $f^* = \sum_{\gamma \in \Gamma} \beta_{\gamma}^* \varphi_{\gamma}$ ).

- learning can be seen as an ill-posed inverse problem and regularization is the theory of choice to restore well-posedness (Girosi and Poggio...)
- in the recent years the genova gang explored the connection between learning and inverse problems in a series of works covering theory, algorithms and applications.

a classical way to avoid overfitting: *penalized empirical risk minimization*

$$\beta_n^\lambda = \underset{\beta}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^n |y_i - f_\beta(x_i)|^2 + \lambda \operatorname{pen}(\beta) \right)$$

different penalizations corresponds to different algorithms:

## examples

- tikhonov/ridge regression:  $\operatorname{pen}(\beta) = \sum_{\gamma \in \Gamma} \beta_\gamma^2$
- basis pursuit/lasso:  $\operatorname{pen}(\beta) = \sum_{\gamma \in \Gamma} |\beta_\gamma|$

we study the regularization scheme defined by

$$\beta_n^\lambda = \underset{\beta}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^n |y_i - f_\beta(x_i)|^2 + \lambda \left( \sum_{\gamma \in \Gamma} w_\gamma |\beta_\gamma| + \varepsilon \sum_{\gamma \in \Gamma} \beta_\gamma^2 \right) \right),$$

(see Zou and Hastie'06)

## Vector notation

$$\left\| \hat{Y} - \Phi_n \beta \right\|_n^2 + \lambda (\|\beta\|_{1,w} + \varepsilon \|\beta\|_2^2)$$

-  $\hat{Y} = (y_1, \dots, y_n)$

-  $\Phi_n$  is  $n \times \Gamma$  (possible infinite dimensional matrix).

# why a combined penalty?

$$\text{pen}(\beta) = \lambda(\|\beta\|_{1,w} + \varepsilon \|\beta\|_2)$$

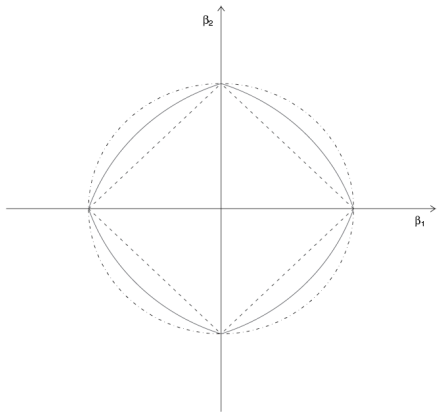
- $\varepsilon > 0$  grouping effect in selection
- $\varepsilon > 0$  strictly convex approximation to basis pursuit
- more stable w.r.t. to noise in the measurements

## linear systems

we (approximately) solve

$$\hat{Y} = \Phi_n \beta$$

where  $\Phi$  is  $n \times \Gamma$ . We look for the minimal  $pen(\beta)$  solution.



questions:

$$\min\left\{\left\|\hat{Y} - \Phi_n\beta\right\|_n^2 + \lambda(\|\beta\|_{1,w} + \varepsilon\|\beta\|_2^2)\right\}$$

Q1: statistical convergence of the algorithm

Q2: solution of the optimization problem

- 1 summability condition:  $\Gamma$  denumerable and, for some  $\kappa > 0$ ,

$$\forall x \in X \quad \sum_{\gamma \in \Gamma} |\varphi_{\gamma}(x)|^2 \leq \kappa.$$

- 2 there exists  $(\beta_{\gamma}^*)_{\gamma \in \Gamma}$  such that

$$\sum_{\gamma \in \Gamma} w_{\gamma} |\beta_{\gamma}^*| < +\infty \quad \text{and} \quad f^*(x) = \sum_{\gamma \in \Gamma} \varphi_{\gamma}(x) \beta_{\gamma}^* \quad x \in X.$$

- 1 summability condition:  $\Gamma$  denumerable and, for some  $\kappa > 0$ ,

$$\forall x \in X \quad \sum_{\gamma \in \Gamma} |\varphi_{\gamma}(x)|^2 \leq \kappa.$$

- 2 there exists  $(\beta_{\gamma}^*)_{\gamma \in \Gamma}$  such that

$$\sum_{\gamma \in \Gamma} w_{\gamma} |\beta_{\gamma}^*| < +\infty \quad \text{and} \quad f^*(x) = \sum_{\gamma \in \Gamma} \varphi_{\gamma}(x) \beta_{\gamma}^* \quad x \in X.$$

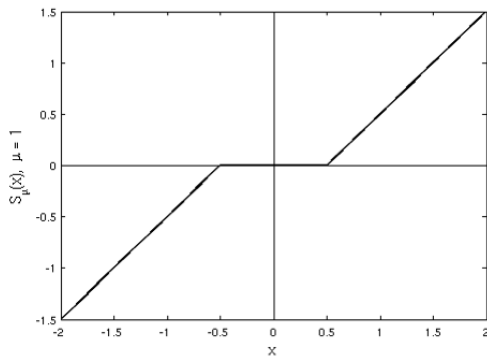
# iterative soft thresholding

let  $\Phi_n$  be the  $n \times \Gamma_\lambda$  matrix (with transpose  $\Phi_n^T$ ).  
We can define an iteration  $\beta^\ell$  converging to  $\beta_n^\lambda$

let  $\beta^0 = \mathbf{0}$ ,  
for  $\ell = 1, 2, \dots$

$$\beta^\ell = \frac{1}{C + \lambda\varepsilon} \mathbf{S}_\lambda \left( (C\mathbf{I} - \Phi_n^T \Phi_n) \beta^{\ell-1} + \Phi_n^T \hat{\mathbf{Y}} \right)$$

# thresholding function



$$S_\mu(x) = \begin{cases} x - \frac{\mu}{2} & \text{if } x > \frac{\mu}{2} \\ 0 & \text{if } |x| \leq \frac{\mu}{2} \\ x + \frac{\mu}{2} & \text{if } x < -\frac{\mu}{2} \end{cases}$$

## generalized solution

$\beta^\varepsilon$  solves

$$\begin{aligned} \min \{ \|\beta\|_{1,w} + \varepsilon \|\beta\|_2 \} \\ \text{s.t. } f^* = \sum_{\gamma \in \Gamma} \varphi_\gamma \beta^* \end{aligned}$$

## elastic net distribution dependent solution

$\beta^\lambda$  solves

$$\min \{ \mathbb{E} [ |f_\beta(x) - y|^2 ] + \lambda \text{pen}(\beta) \}$$

we also let  $f^\lambda = \sum_{\gamma \in \Gamma} \varphi_\gamma \beta_\gamma^\lambda$

Error decomposition for fixed  $\lambda > 0$

$$\left\| \beta_n^\lambda - \beta^\varepsilon \right\|_2 \leq \underbrace{\left\| \beta_n^\lambda - \beta^\lambda \right\|_2}_{\text{sample error}} + \underbrace{\left\| \beta^\lambda - \beta^\varepsilon \right\|_2}_{\text{approximation error}},$$

- under some assumptions on the noise with probability greater than  $1 - 4e^{-\delta}$

$$\left\| \beta_n^\lambda - \beta^\lambda \right\|_2 \leq \left\| \beta^\lambda - \beta \right\|_2 \left( \frac{C\sqrt{\delta}}{\sqrt{n\lambda}} \right) + \left( \frac{C\sqrt{\delta}}{\sqrt{n\lambda}} \right)$$

- The approximation error satisfies

$$\lim_{\lambda \rightarrow 0} \left\| \beta^\lambda - \beta^\varepsilon \right\|_2 \rightarrow 0$$

if we choose  $\lambda_n$  s.t.  $\lambda_n \sqrt{n} \rightarrow \infty$ , when  $n \rightarrow \infty$  then

$$\mathbb{E} \left[ \left\| \beta_n^{\lambda_n} - \beta^\varepsilon \right\|_2 \right] \rightarrow 0,$$

moreover we also have

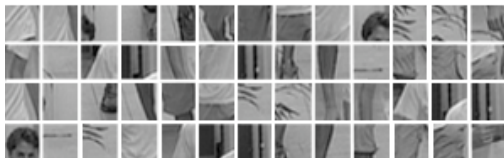
$$\mathbb{E} \left[ \mathcal{E}(f_n^{\lambda_n}) - \mathcal{E}(f^*) \right] \rightarrow 0$$

- many algorithms for sparse selection, (Mallat et al. - Gilbert and Tropp OMP, Candes and Tao 06 - Dantzig estimator, Donoho et al. - Basis Pursuit, Fugueredo and Nowak - Projected Gradient, Freund and Shapiro- Boosting...)
- many theoretical results on L2 regularization (Smale and Zhou 05, Caponnetto and De Vito 05, Bauer et al. 06...).
- Recently many results for sparsity based scheme. Mostly in different settings - fixed design regression, signal processing, linear coding - (Donoho '85...Candes and Tao 05... Daubachies et al. 05...)
- fewer results in the context of learning (Barron et al. 06, Bunuea et al. 06, Koltchinskii 07...)

# application I: face detection

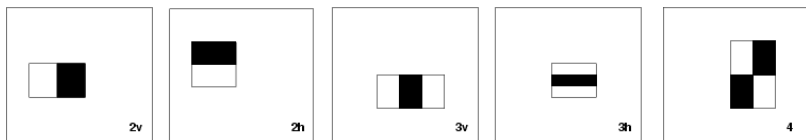
(Destrero, De Mol, Odone, Verri 07)

face detection integrated in a monitoring system in our department

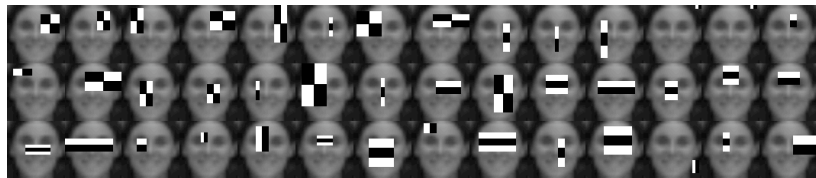


data:image size 20x20, 2000 + 2000 training, 1000 + 1000 validation, 2000 + 2000 test.

overcomplete dictionary of rectangular features capturing the local geometry of faces (Viola and Jones, 2004),

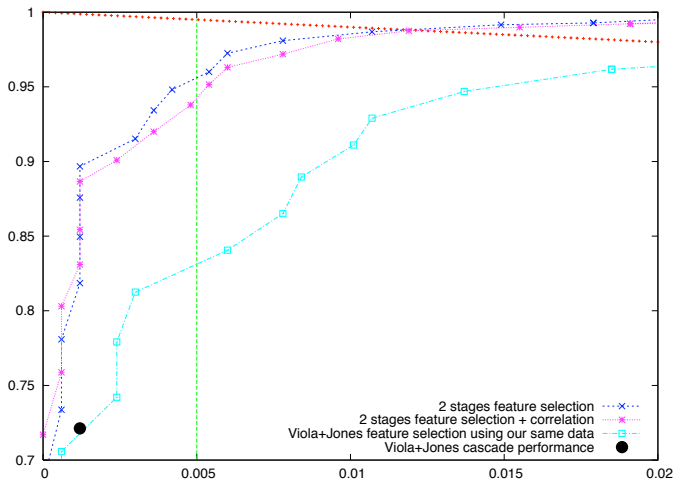


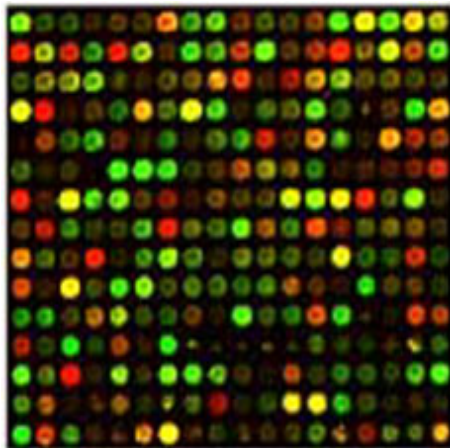
- features computed at all locations, scales, aspect ratios: roughly 64000 per image.
- highly correlated features



42 features extracted by a 2 stage selection scheme.

# results





in micro-array classification the number of samples is much smaller than the number of genes expressions

(Mosci, De Mol, Traskine, Verri 07)

Algorithms were tested on three datasets:

- leukemia (patients 72 (38-34) , genes 7129)
- lung cancer (patients 181 (91-90), genes 12533)
- prostate cancer (patients 102 (51-51), genes 12533)

$\lambda = 0.07$ $\lambda \varepsilon$	test error (test set size: 90)	# of selected genes	intersection w. genes selected for bigger $\varepsilon$
0	0	22	100%
0.001	0	28	100%
0.0025	1	37	100%
0.005	1	54	100%
0.01	1	80	99%
0.1	1	247	96%
1	1	743	—

Previous: 8 genes with 91-98 % correct classification on test  
(Gordon et al. 02)

$\lambda = 0.06$ $\lambda\varepsilon$	test error (test set size: 51)	# of selected genes	intersection w. genes selected for bigger $\varepsilon$
0	5	19	95%
0.001	5	20	100%
0.0025	5	25	100%
0.005	4	31	97%
0.01	5	40	98%
0.1	6	85	94%
1	5	121	—

Previous: 5-8 genes with 82,9-95,7 % correct classification on test using ranking and K-NN (Singh et al 2002).

- approximation properties and connections with work in signal processing
- design good (data-dependent?) dictionary