Regularization frontier in machine learning

G. Gasso

LITIS *EA 4108*
INSA de Rouen, France

ECML PKDD 2008

September 19th, 2008

Antwerp, Belgium

Roadmap

1. Introduction
   - Framework
   - Model selection
2. Regularization path and pareto frontier
3. Efficient regularization path running
   - Piecewise linear regularization path
4. Two examples of regularization path
   - Lasso path
   - SVM path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Framework

- Set of data $\mathcal{D} = \{x_i, y_i\}_{i=1,\cdots,n}$ with $(x, y) \in \mathcal{X} \times \mathcal{Y}$
  $(X, Y) \sim P_{X,Y}$ with $P_{X,Y}$ the unknown joint distribution

- Supervised learning
  - Binary classification $\mathcal{Y} = \{-1, +1\}$
  - Regression $\mathcal{Y} = \mathbb{R}$
- Task : find a predictive model f

$$\begin{aligned} f : \quad \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto \hat{y} = f(x) \end{aligned}$$

- f belongs to an hypothesis space $\mathcal{H}$

Framework (ct'd)

- A non-negative loss function ℓ
- Expected risk minimization $f^\star = \text{argmin}_{f \in \mathcal{H}} E_{X,Y}\left(\ell(f(X), Y)\right)$
- Empirical loss minimization

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}}\, L(f)$$

with $L(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

- To avoid overtraining, some constraints (smoothness, sparsity, robustness, ...) are enforced on f by using a Regularizer or penalty term $P(f)$
- Regularized optimization problem

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}}\, L(f) + \lambda\, P(f)$$

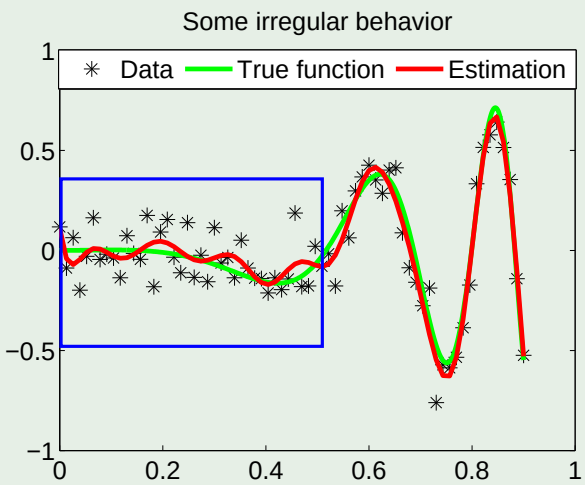
$\lambda \in \mathbb{R}^+$ is a trade-off or regularization parameter

Tuning of λ

Identify the appropriate value $\lambda^\star$ associated to the best solution $\hat{f}^\star$

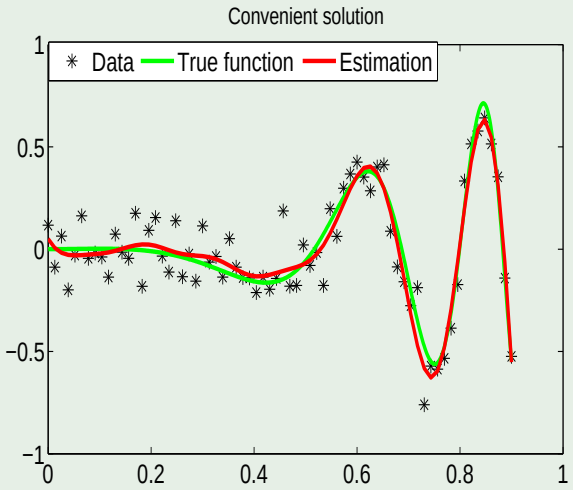
Illustration : non linear ridge regression

Overfitting



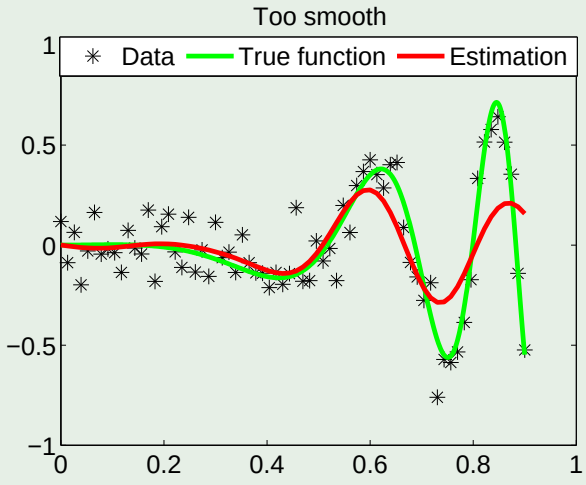
Small λ

Convenient solution



Mid value of λ

Underfitting



High λ

Determination of λ

- Compute the decision function $\hat{f}_\lambda$ for different values of λ
- Select the best solution according to some generalization performance

Model selection

Determination of λ

- Compute the decision function $\hat{f}_\lambda$ for different values of λ
- Select the best solution according to some generalization performance

Two approaches

Model selection

Determination of λ

- Compute the decision function $\hat{f}_\lambda$ for different values of λ
- Select the best solution according to some generalization performance

Two approaches

1. Grid search over predefined set $\{\lambda_1, \ldots, \lambda_K\}$

- Values specified by the user
- Retained solution $\hat{f}^\star(x)$ depends highly on the grid resolution

Model selection

Determination of λ

- Compute the decision function $\hat{f}_\lambda$ for different values of λ
- Select the best solution according to some generalization performance

Two approaches

1. Grid search over predefined set $\{\lambda_1, \ldots, \lambda_K\}$
2. Compute the regularization path

- No values specified by the user
- Find automatically all solutions $\hat{f}_\lambda(x)$

Regularization path

The set of all solutions $\hat{f}_\lambda(x)$ i.e. $\mathcal{R} = \left\{ \hat{f}_\lambda(x) \mid \lambda \in [0, \infty) \right\}$

Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Linear ridge regression

- Model : $f(x) = x^\top \beta$ with $\beta \in \mathbb{R}^d$
- Problem :

$$\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$$

- Solution : $\hat{\beta}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{y}$

I : identity matrix

Regularization path

- $\mathcal{R} = \left\{\hat{\beta}(\lambda) \mid \lambda \in [0, \infty)\right\}$
- $\lambda = 0$, $\hat{\beta}_{LS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$ (least squares solution)
- $\lambda \to \infty$, $\hat{\beta} = \mathbf{0}$

1D linear regression path

$$\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (x_i \beta - y_i)^2 + \lambda \beta^2$$

The Loss L as a function of the regularizer P

$$\begin{cases} L(\beta) = \sum_{i=1}^{n} (x_i \beta - y_i)^2 \\ P(\beta) = \beta^2 \end{cases}$$

It holds that

$$\begin{cases} L(P) = aP \pm b\sqrt{P} + c \\ a, b \text{ and } c \in \mathbb{R} \end{cases}$$

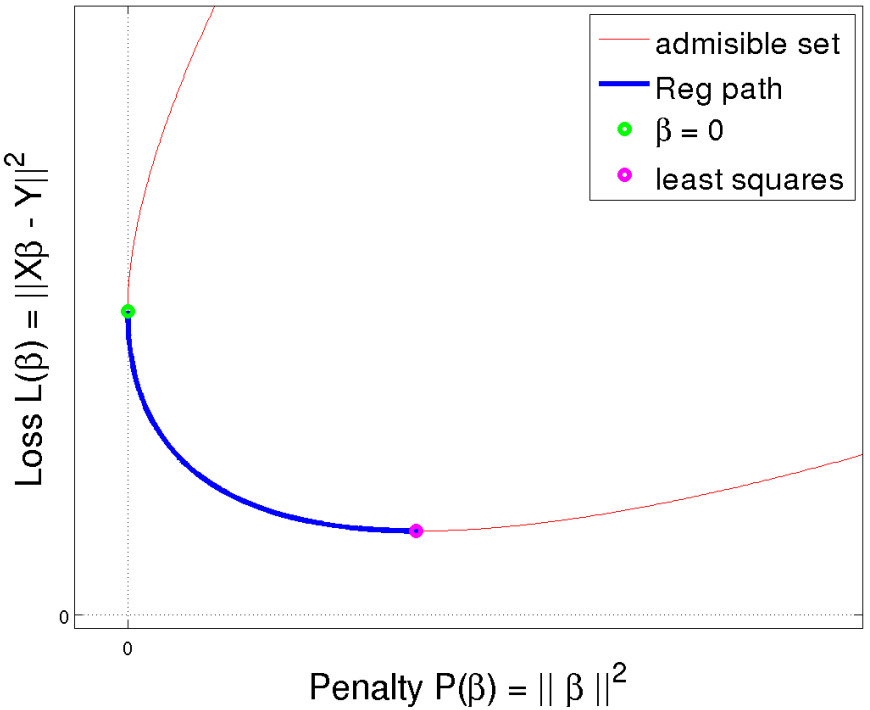


Fig.: regularization path as a function of L and P

Notion of Dominance and Pareto frontier

$$\begin{cases} L(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ P(\beta) = \|\beta\|^2 \end{cases}$$

Dominance

A vector β_1 dominates another vector β_2 if $L(\beta_1) \leq L(\beta_2)$ and $P(\beta_1) \leq P(\beta_2)$

Pareto frontier

Pareto frontier is the set of all non dominated solutions

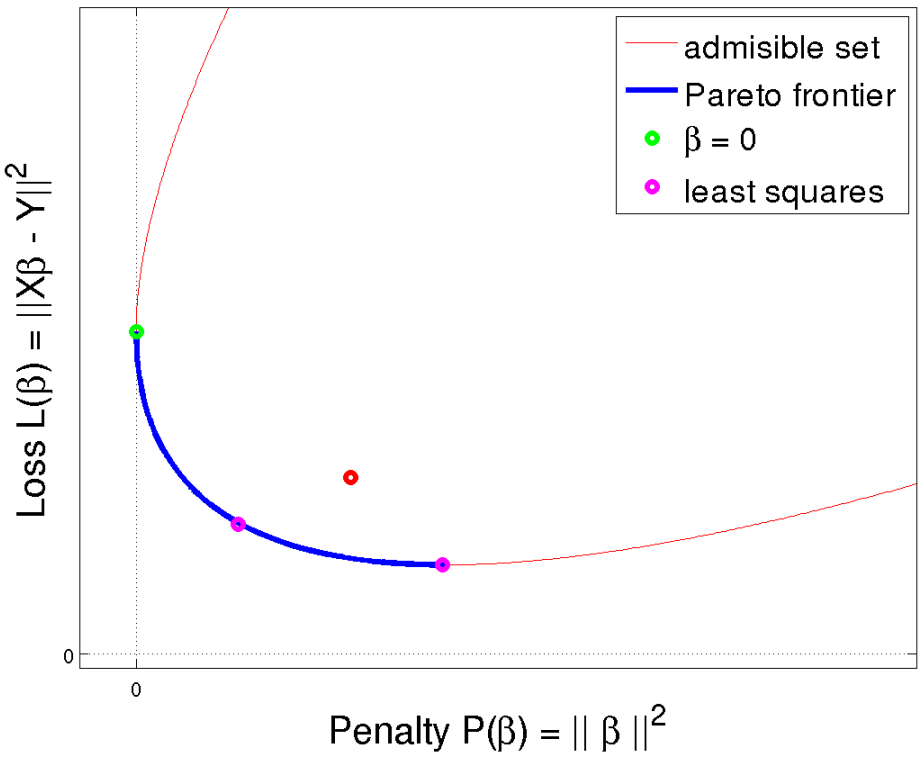


Fig.: dominated point (red), non dominated point (purple) and Pareto frontier (blue).

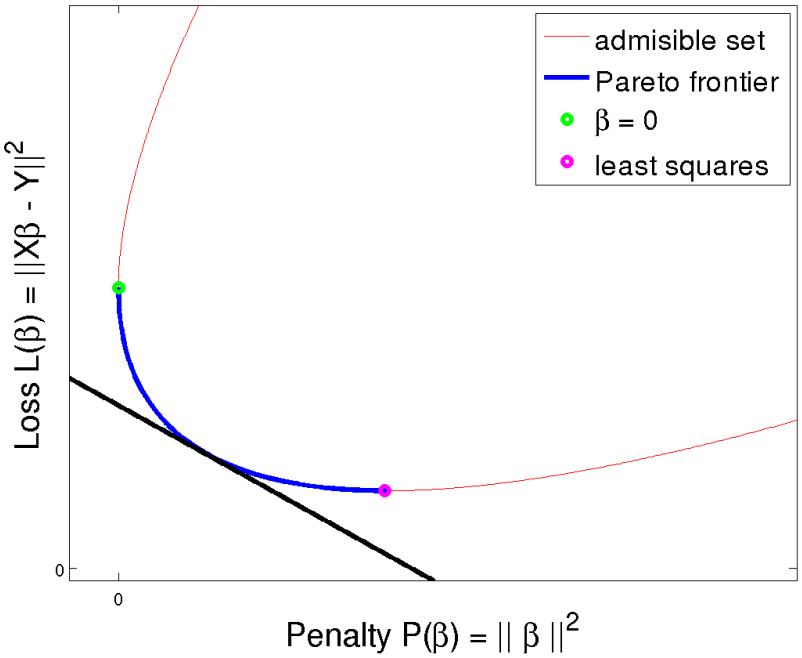
Pareto frontier $\Leftrightarrow$ Reg. path

It works for CONVEX criteria !

Formulation 1 : Lagrangian

(linear combination of L and P)

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$$



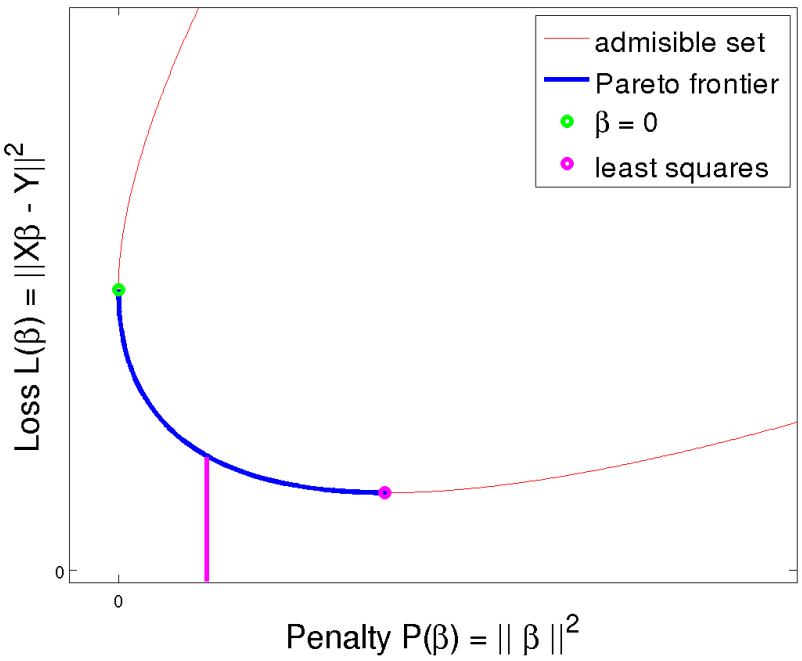
It works for CONVEX criteria !

Formulation 1

$$\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2$$

Formulation 2

$$\begin{cases} \min_{\beta} \|y - X\beta\|^2 \\ \text{s.t. } \|\beta\|^2 \leq C \end{cases}$$



It works for CONVEX criteria !

Formulation 1

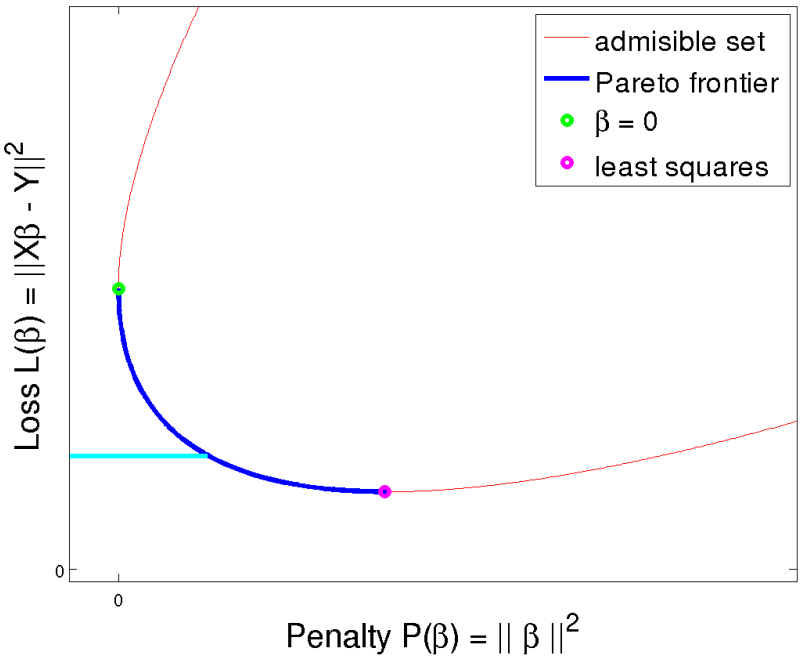
$$\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2$$

Formulation 2

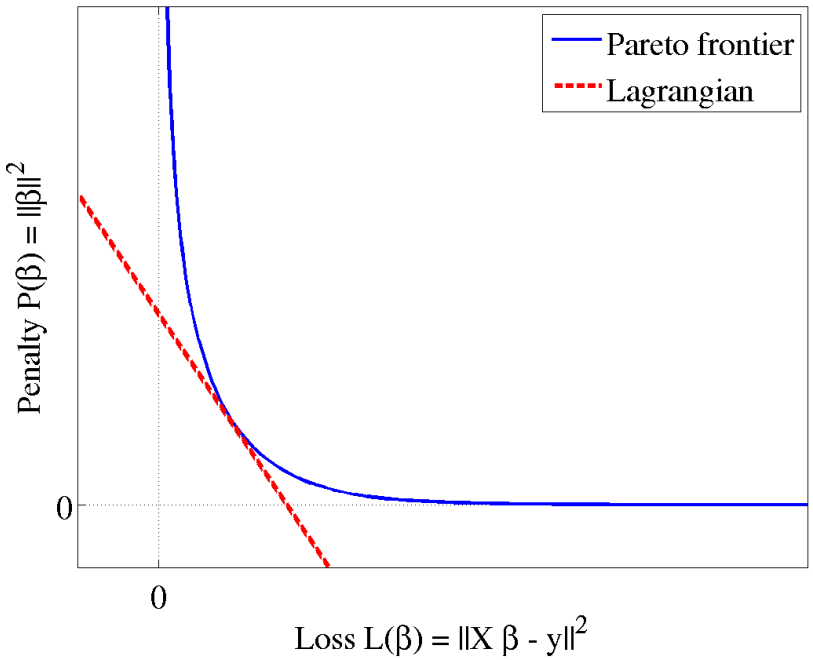
$$\begin{cases} \min_{\beta} \|y - X\beta\|^2 \\ \text{s.t. } \|\beta\|^2 \leq C \end{cases}$$

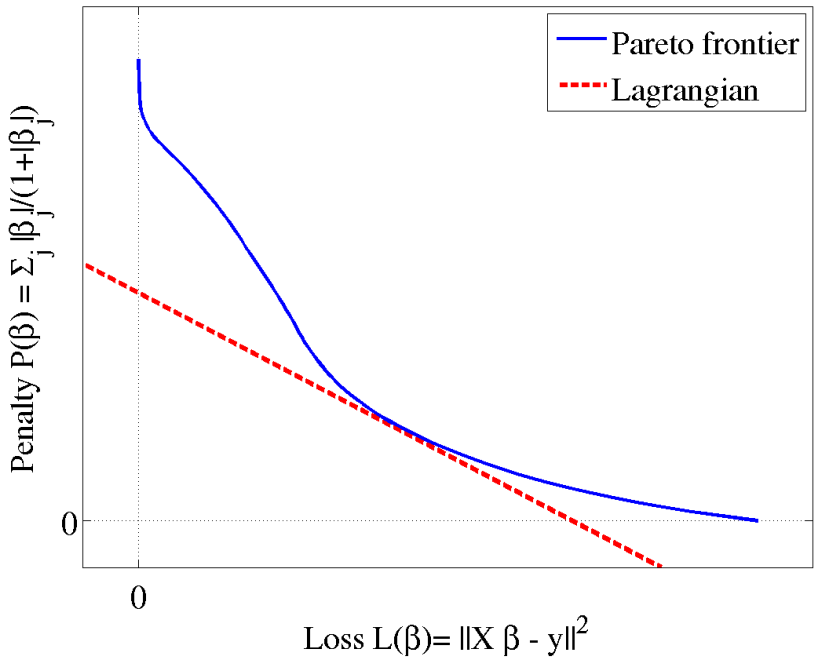
Formulation 3

$$\begin{cases} \min_{\beta} \|\beta\|^2 \\ \text{s.t. } \|y - X\beta\|^2 \leq C' \end{cases}$$



The importance of convexity





Non convex case

The 3 formulations are not equivalent

so far...

- learning is a multi objective problem
- the regularization path is the Pareto frontier
- beware the non convex case
- it works for more than 2 criteria

What for ?

To tune (*efficiently*) the regularization parameter λ

Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Tuning the regularization parameter λ

Ridge regression example $\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$

Tuning the regularization parameter λ

Ridge regression example $\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$

- Grid Search

  for each $\lambda_1 < \lambda_2 < ... < \lambda_t < ... < \lambda_K$
  compute $\beta_t = (\mathbf{X}^\top \mathbf{X} + \lambda_t I)^{-1} \mathbf{X}^\top \mathbf{y}\,, \quad t = 1, \cdots, K$

  $\mathcal{O}\left(Kd^3\right)$

Tuning the regularization parameter λ

Ridge regression example $\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$

- Grid Search

  for each $\lambda_1 < \lambda_2 < ... < \lambda_t < ... < \lambda_K$
  compute $\beta_t = (\mathbf{X}^\top \mathbf{X} + \lambda_t I)^{-1} \mathbf{X}^\top \mathbf{y}\,, \quad t = 1, \cdots, K$

  $\mathcal{O}\left(Kd^3\right)$

- Warm start

  $\beta_t = \Phi(\beta_{t-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\left(K\ell d^2\right)$

Tuning the regularization parameter λ

Ridge regression example $\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$

- Grid Search

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_t < \ldots < \lambda_K$
  compute $\beta_t = (\mathbf{X}^\top \mathbf{X} + \lambda_t I)^{-1} \mathbf{X}^\top \mathbf{y}\,, \quad t = 1, \cdots, K$

  $\mathcal{O}\left(Kd^3\right)$

- Warm start

  $\beta_t = \Phi(\beta_{t-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\left(K\ell d^2\right)$

- Warm start + prediction step

  $\beta_t^{(p)} = \beta_{t-1} + \rho \nabla_\beta (L(\beta_{t-1}) + \lambda_t P(\beta_{t-1}))$ (prediction step)

  $\beta_t = \Phi(\beta_t^{(p)})$ (correction step using conjugate gradient)

  $\mathcal{O}\left(K\ell' d^2\right)$

Tuning the regularization parameter λ

Ridge regression example $\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$

- Grid Search

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_t < \ldots < \lambda_K$
  compute $\beta_t = (\mathbf{X}^\top \mathbf{X} + \lambda_t I)^{-1} \mathbf{X}^\top \mathbf{y}\,, \quad t = 1, \cdots, K$

  $\mathcal{O}(Kd^3)$

- Warm start

  $\beta_t = \Phi(\beta_{t-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}(K\ell d^2)$

- Warm start + prediction step

  $\beta_t^{(p)} = \beta_{t-1} + \rho \nabla_\beta (L(\beta_{t-1}) + \lambda_t P(\beta_{t-1}))$ (prediction step)
  $\beta_t = \Phi(\beta_t^{(p)})$ (correction step using conjugate gradient)

  $\mathcal{O}(K\ell' d^2)$

- Use only the prediction step !

  $\beta_t = \beta_{t-1} + \lambda_t \Psi(\beta_{t-1})$ (prediction step)
  to do so the regularization path has to be piecewise linear

  $\mathcal{O}(Kd^2)$

Piecewise linearity conditions

$$\min_{\beta \in \mathbb{R}^d} L(\beta) + \lambda P(\beta)$$

How to choose L and P to get linear reg. path ?

Solution path is linear $\Leftrightarrow$ one cost is piecewise quadratic and the other one piecewise linear

convex case [Rosset & Zhu, 07]

Piecewise linearity can be interpreted as

$$\lim_{\varepsilon \to 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = \text{constant}$$

Piecewise linearity conditions : proof

1. Optimality conditions

$$\text{For} \quad \lambda \qquad \longrightarrow \qquad \nabla L(\beta(\lambda)) + \lambda \nabla P(\beta(\lambda)) = 0$$

$$\text{For} \quad \lambda + \varepsilon \qquad \longrightarrow \qquad \nabla L(\beta(\lambda+\varepsilon)) + (\lambda+\varepsilon) \nabla P(\beta(\lambda+\varepsilon)) = 0$$

2. Use first order Taylor expansion around $\beta(\lambda)$

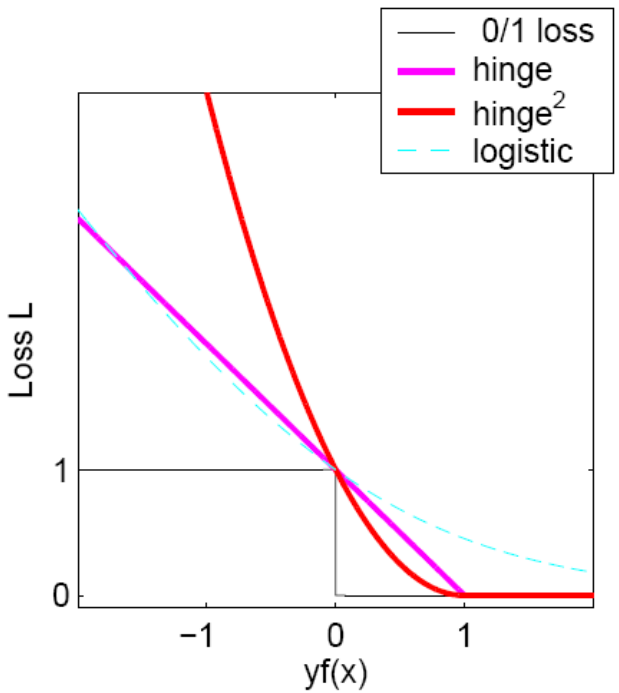
$$\begin{aligned} &\nabla L(\beta(\lambda)) + \nabla^2 L(\beta(\lambda))[\beta(\lambda+\varepsilon) - \beta(\lambda)] \\ &+\lambda \nabla P(\beta(\lambda)) + \lambda \nabla^2 P(\beta(\lambda))[\beta(\lambda+\varepsilon) - \beta(\lambda)] + \varepsilon \nabla P(\beta(\lambda)) + O(\varepsilon^2) = 0 \end{aligned}$$

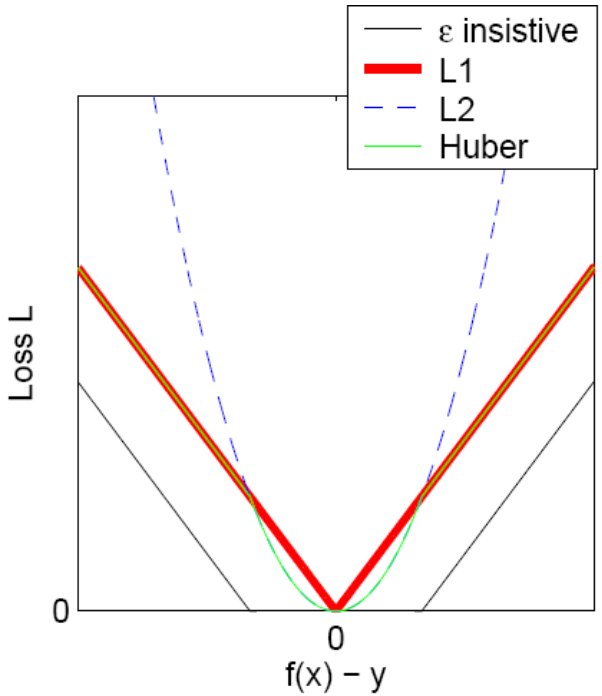
3. Variation of β according to λ

$$\lim_{\varepsilon \longrightarrow 0} \frac{\beta(\lambda+\varepsilon) - \beta(\lambda)}{\varepsilon} = -\left[\nabla^2 L(\beta(\lambda)) + \lambda \nabla^2 P(\beta(\lambda))\right]^{-1} \nabla P(\beta(\lambda))$$

$$\nabla^2 L(\beta(\lambda)) = \text{constant} \quad \text{and} \quad \nabla^2 P(\beta(\lambda)) = 0$$

Examples of Loss and Penalty





L	P	*regression*	*classification*	*clustering*
L_2	L_1	Lasso/LARS	L1 L2 SVM	
L_1	L_2	SVR	SVM	OneClass SVM
L_1	L_1	L1 least absolute deviation	L1 SVM	

Tab.: example of piecewise linear regularization path algorithms.

$$P: \quad L_p = \sum_{j=1}^{d} |\beta_j|^p \qquad L: \quad L_p : |f(x) - y|^p \quad \text{hinge } (yf(x) - 1)_+^p$$

ε-insensitive
$$\begin{cases} 0 & \text{if } |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise} \end{cases}$$

Huber's loss :
$$\begin{cases} |f(x) - y|^2 & \text{if } |f(x) - y| < t \\ 2t|f(x) - y| - t^2 & \text{otherwise} \end{cases}$$

Piecewise regularization path

- the problem

$$\min_{\beta \in \mathbb{R}^d} L(\beta) + \lambda P(\beta) \quad \Leftrightarrow \quad \{\beta(\lambda) \mid \lambda \in [0, \infty]\}$$

- L and P are convex
- efficient computation

$$\Longrightarrow \text{piecewise linearity}$$

$$\beta_{t+1} = \beta_t + (\lambda_{t+1} - \lambda_t)\mathbf{w}$$

- piecewise linearity

$$\Longrightarrow \text{either } L \text{ or } P \text{ is } L_1 \text{ type}$$

An old result revisited

- Portfolio management (Markovitz, 1952)
  - Gain *vs.* risk
$$\begin{cases} \min_{\beta} & \frac{1}{2}\beta^\top Q\beta \\ \text{with} & \mathbf{e}^\top\beta = C \end{cases}$$
  - *efficiency frontier* : piecewise linearity (*Critical path Algo.*)

- Sensitivity analysis (Heller, 1954) I. HELLER (1954) Sensitivity analysis in linear programming. L.R.P. Seminar, The George Washington University, Logistics Research Project, January 1954.
$$\begin{cases} \min_{\beta} & \frac{1}{2}\beta^\top Q\beta + (\mathbf{c} + \lambda\ \Delta\mathbf{c})^\top\beta \\ \text{avec} & A\beta = \mathbf{b} + \mu\ \Delta\mathbf{b} \end{cases}$$

- Parametric programming (Gal 1968)
  - Parametric Linear Programming is piecewise linear
  - PQP piecewise quadratic
  - Multiparametric programming...

Roadmap

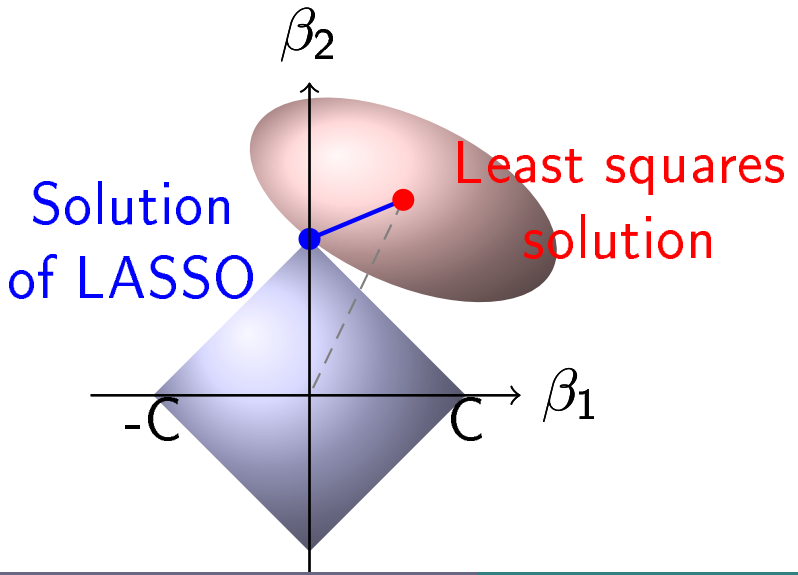
1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Lasso (Basis pursuit) problem

$$\min_{\beta \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^{d} |\beta_j| \iff \begin{cases} \min_{\beta \in \mathbb{R}^d} & \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ \text{with} & \sum_{i=1}^{d} |\beta_j| \leq C \end{cases}$$

- Variables $x_j = \mathbf{X}(:, j)$ for $j = 1, \cdots, d$
- Assume all x_j and $\mathbf{y}$ are centered and normalized

Illustration on two dimensional example



small C leads to $\beta_1 = \beta_2 = 0$

high C produces the least squares solution

Lasso regularization path

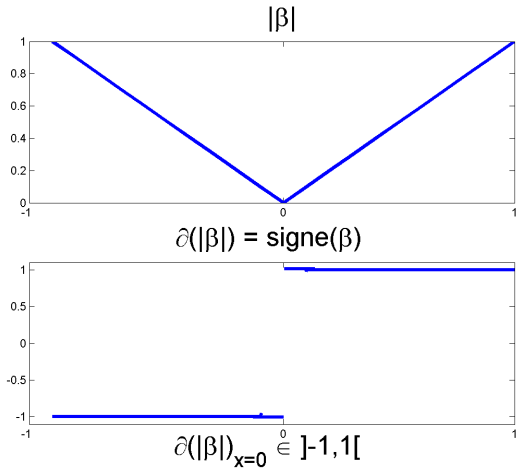
$$\min_{\beta} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^{d} |\beta_j|$$

Optimality condition for variable x_j

$$\underbrace{-x_j^\top (\mathbf{X}\beta - \mathbf{y})}_{\textit{correlation}} + \lambda \, \partial(|\beta_j|) = 0$$

Subdifferential

$$\partial(|\beta_j|) = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \alpha_j \in [-1, 1] & \text{if } \beta_j = 0 \end{cases}$$



Lasso regularization path

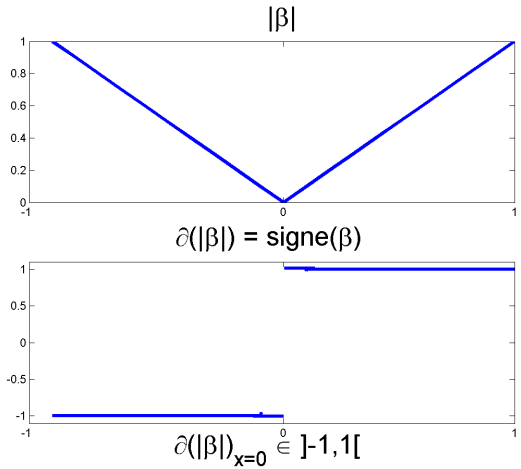
$$\min_{\beta} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^{d} |\beta_j|$$

Optimality condition for variable x_j

$$\underbrace{-x_j^\top(\mathbf{X}\beta - \mathbf{y})}_{\textit{correlation}} + \lambda\, \partial(|\beta_j|) = 0$$

Subdifferential

$$\partial(|\beta_j|) = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \alpha_j \in [-1, 1] & \text{if } \beta_j = 0 \end{cases}$$



Active set : $I_\beta = \{\beta_j \mid \beta_j \neq 0\}$

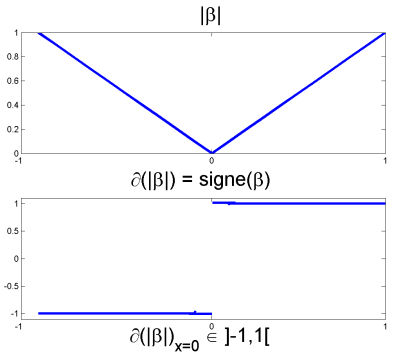
$$\underbrace{|x_j^\top(\mathbf{X}\beta - \mathbf{y})|}_{\textit{correlation}} = \lambda, \quad \beta_j \in I_\beta$$

Inactive set $I_0 = \{\beta_j \mid \beta_j = 0\}$

$$\underbrace{|x_j^\top(\mathbf{X}\beta - \mathbf{y})|}_{\textit{correlation}} \leq \lambda, \quad \beta_j \in I_0$$

Lasso regularization path

$$\min_\beta \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{d} |\beta_j|$$



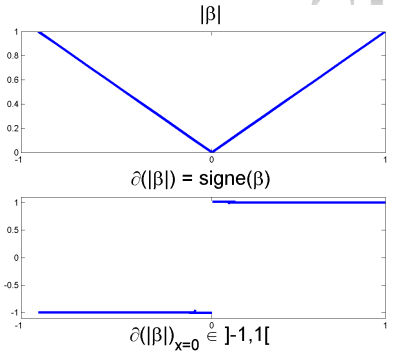
- Let $\boldsymbol{\beta}_\beta = \boldsymbol{\beta}(I_\beta)$ and $\mathbf{X}_\beta = \mathbf{X}_\beta(:, I_\beta)$
- optimality conditions become

$$-\mathbf{X}_\beta^\top(\mathbf{X}_\beta\boldsymbol{\beta}_\beta - \mathbf{y}) + \lambda \operatorname{sign}(\boldsymbol{\beta}_\beta) = 0$$

Lasso regularization path

- Let $\beta_\beta = \beta(I_\beta)$ and $\mathbf{X}_\beta = \mathbf{X}_\beta(:, I_\beta)$
- optimality conditions become

$$-\mathbf{X}_\beta^\top(\mathbf{X}_\beta\beta_\beta - \mathbf{y}) + \lambda\, \text{sign}(\beta_\beta) = 0$$



- For λ_t, assume the solution β_β^t and the corresponding set I_β^t
- Assume $\lambda = \lambda_t + \gamma$ such that I_β^t and $\text{sign}(\beta_\beta^t)$ remain unchanged

$$\begin{array}{rcl} \mathbf{X}_\beta^\top(\mathbf{X}_\beta\beta_\beta^t - \mathbf{y}) & = & \lambda_t\, \text{sign}(\beta_\beta) \\ \mathbf{X}_\beta^\top(\mathbf{X}_\beta\beta_\beta - \mathbf{y}) & = & \lambda\, \text{sign}(\beta_\beta) \end{array}$$

$$\mathbf{X}_\beta^\top\mathbf{X}_\beta(\beta_\beta - \beta_\beta^t) = (\lambda - \lambda_t)\, \text{sign}(\beta_\beta)$$

$$\beta_\beta = \beta_\beta^t + (\lambda - \lambda_t)\mathbf{w} = \beta_\beta^t + \gamma\, \mathbf{w}$$

Descent direction $\mathbf{w} = (\mathbf{X}_\beta^\top\mathbf{X}_\beta)^{-1}\text{sign}(\beta_\beta)$

Lasso regularization path

$$\boldsymbol{\beta}_\beta \quad = \quad \boldsymbol{\beta}^t_\beta + \gamma\, \mathbf{w}$$

The linear variation holds until the set I^t_β changes $\Longrightarrow$ detect events

Event detection

- $\beta_\ell \in I_\beta$ moves to I_0

  Compute the step size γ such as $0 = \beta^t_\ell + \gamma\, \mathbf{w}_j$

- $\beta_j \in I_0$ moves to I_β

  Recall $\quad |\mathrm{x}_\ell^\top (\mathbf{X}_\beta \boldsymbol{\beta}_\beta - \mathbf{y})| = \lambda, \quad \beta_\ell \in I_\beta \quad$ and $\quad |\mathrm{x}_j^\top (\mathbf{X}_\beta \boldsymbol{\beta}_\beta - \mathbf{y})| \leq \lambda, \quad \beta_j \in I_0$

  Compute γ to obtain the correlation $|\mathrm{x}_j^\top \left(\mathbf{X}_\beta \boldsymbol{\beta}_\beta - \mathbf{y}\right)| = \lambda_t + \gamma$

  Choose β_j as the most correlated variable to the residual i.e.
  $j = \mathrm{argmax}_{j \in I_0} \ |\mathrm{x}_j^\top (\mathbf{X}_\beta \boldsymbol{\beta}^t_\beta - \mathbf{y})|$

Algorithm

Algorithm 1 Lasso solution path

Set $t = 0$, $\beta^0 = 0$, $I_\beta = \emptyset$ and $I_0 = \{1, \cdots, d\}$

Find β_j to add to I_β : $j = \text{argmax } |\mathbf{x}_j^\top \mathbf{y}|, \quad j \in I_0$ (max of correlation)

repeat

Compute the descent direction $\mathbf{w}$

Compute the step size γ

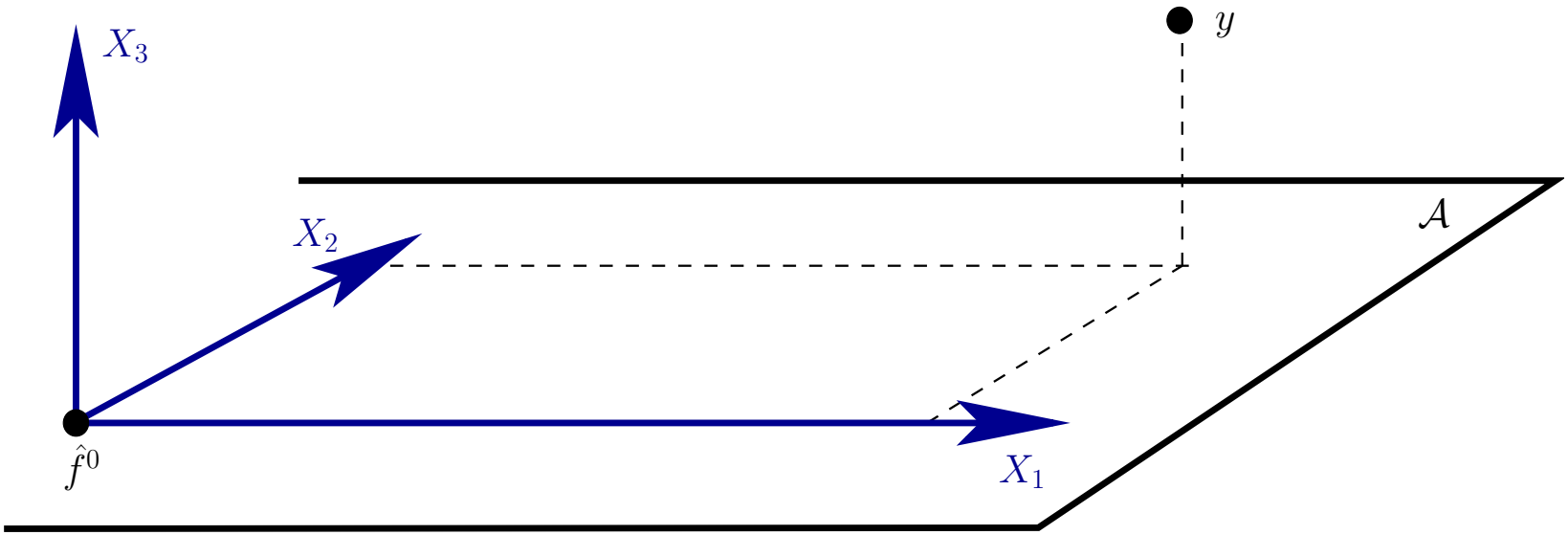
Update the sets I_β and I_0 according to the event detected ($I_0 \rightarrow I_\beta$ or $I_\beta \rightarrow I_0$)

$t = t + 1$

until termination

Interpretation of Lasso path (V. Guigue)

the solution in the X space

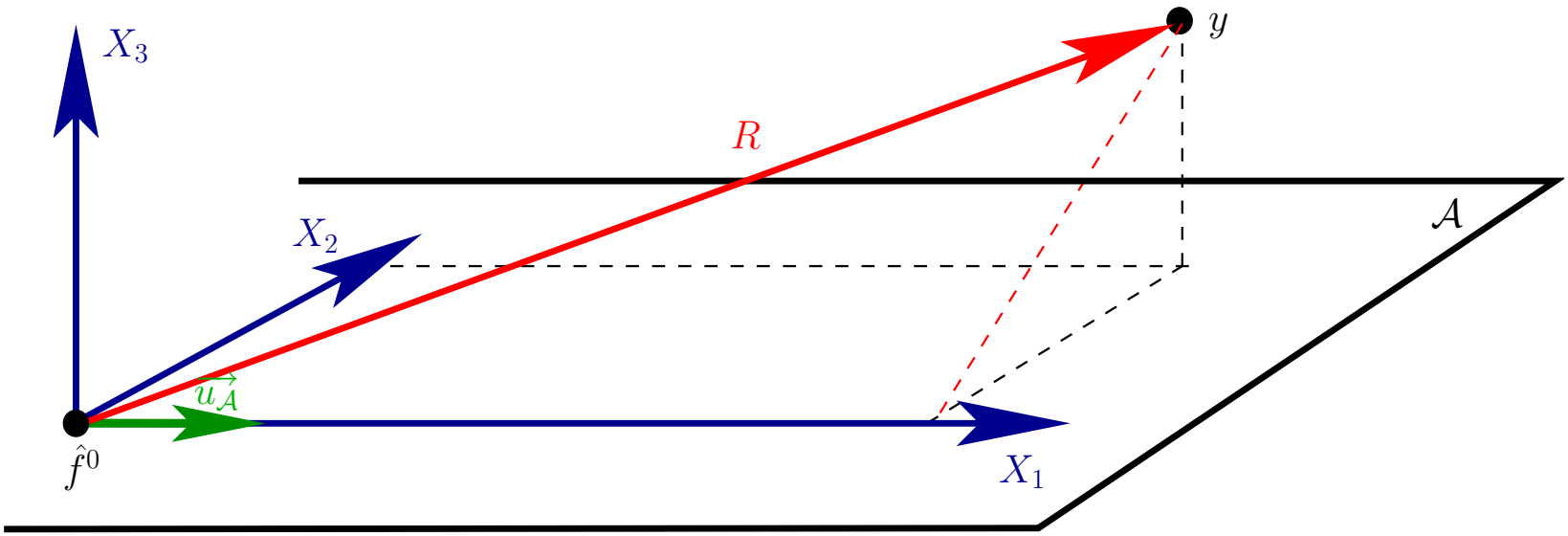


starting point : all the β are set to 0

residual : $\mathbf{R} = X\boldsymbol{\beta} - \mathbf{y} = \mathbf{y}$

Interpretation of Lasso path (V. Guigue)

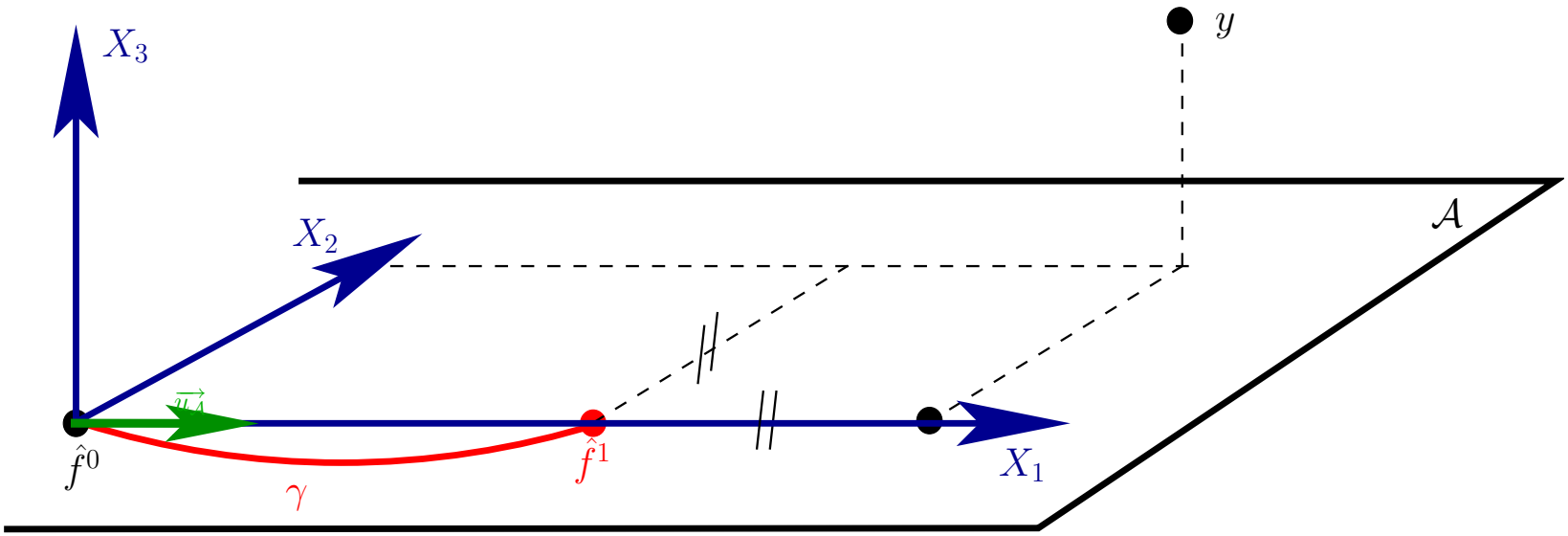
the solution in the X space



projection of the residual error on the active variable

Interpretation of Lasso path (V. Guigue)

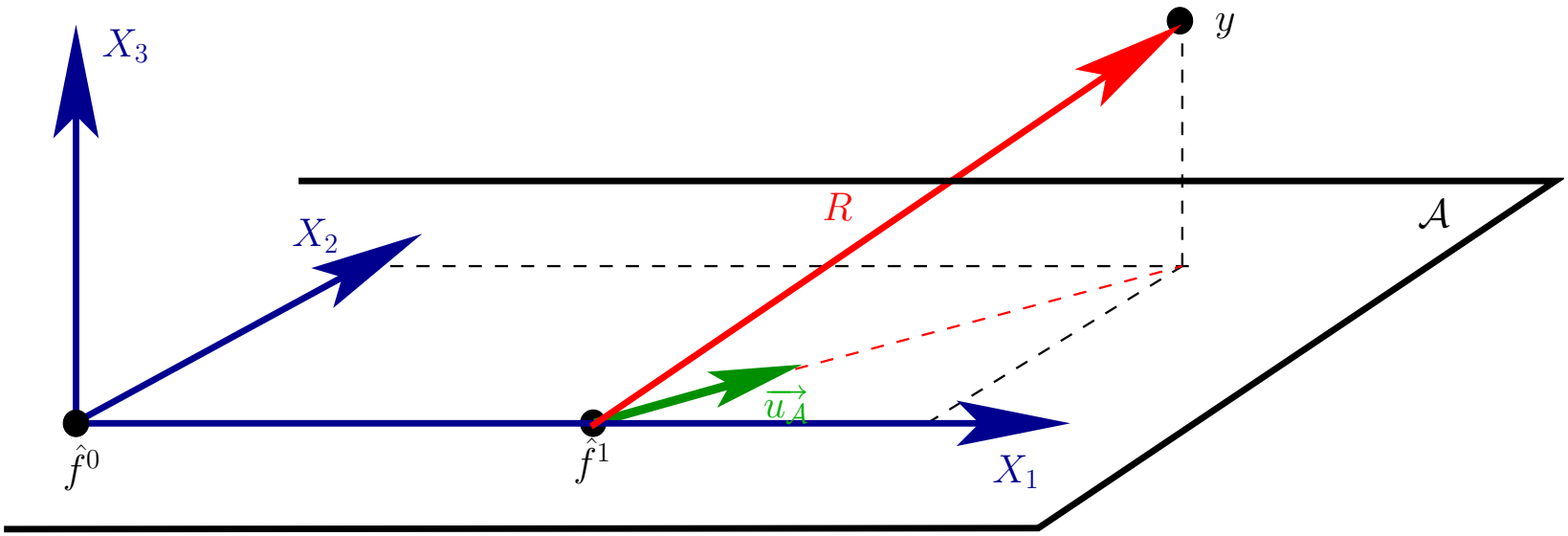
the solution in the X space



stepsize computation, same correlation of residual errors

Interpretation of Lasso path (V. Guigue)

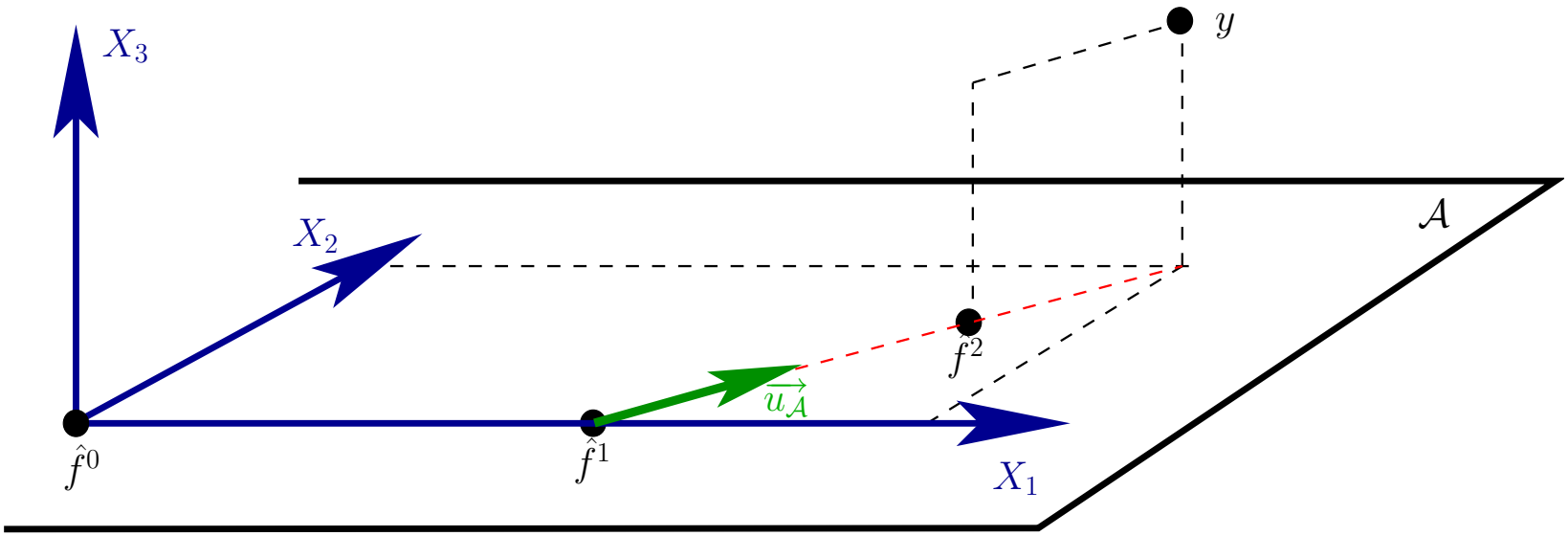
the solution in the X space



projection of the residual error on the active variable

Interpretation of Lasso path (V. Guigue)

the solution in the X space



stepsize computation

Illustration of Lasso regularization path

Example (provided by A. Rakotomamonjy)

- Diabetes data set : 10 variables, 442 observations

Parameters β Output

Linear SVM

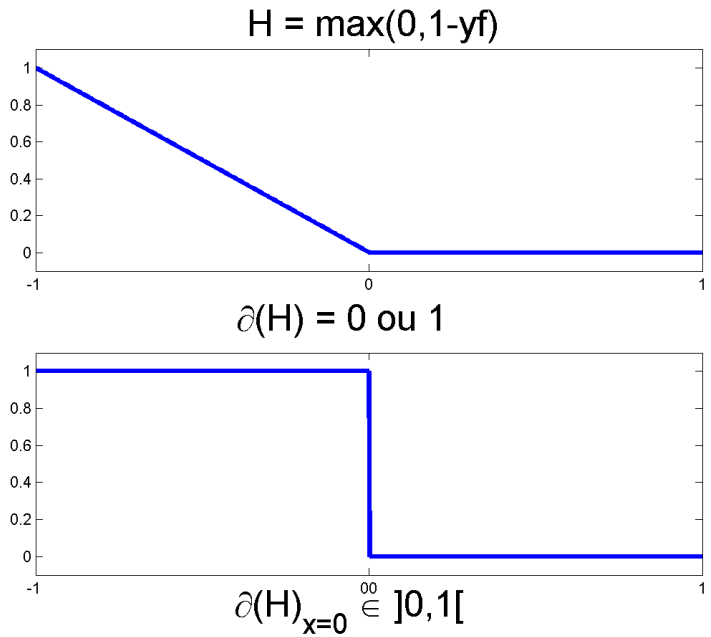
$$\text{Model :} \quad f(x) = \langle \omega, x \rangle \quad \text{(to easy the presentation)}$$

$$\text{Problem :} \quad \min_\omega \sum_{i=1}^{n} \underbrace{\max(1 - y_i f(x_i), 0)}_{\text{hinge loss}} + \frac{\lambda}{2} \|\omega\|^2$$

Optimality condition

$$-\sum_{i=1}^{n} \alpha_i y_i x_i^\top + \lambda \omega = 0$$

$$\begin{cases} \alpha_i = 1 & \text{if } y_i f(x_i) < 1 \\ \alpha_i = 0 & \text{if } y_i f(x_i) > 1 \\ \alpha_i \in [0,1] & \text{if } y_i f(x_i) = 1 \end{cases}$$



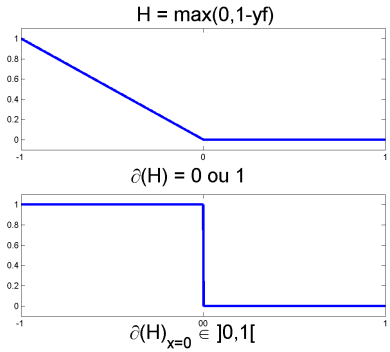
Sets

$$I_0 = \{x_i \mid y_i f(x_i) > 1\}, \quad I_1 = \{x_i \mid y_i f(x_i) < 1\}, \quad I_\alpha = \{x_i \mid y_i f(x_i) = 1\}$$

Linear SVM path

Optimality condition

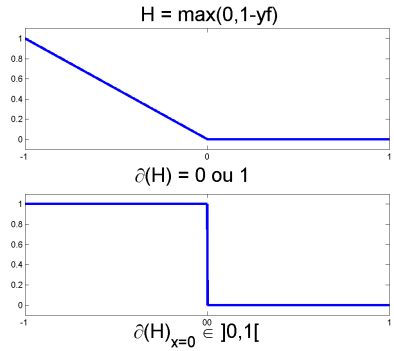
$$\sum_{I_\alpha} \alpha_i y_i x_i^\top + \sum_{I_1} y_i x_i^\top = \lambda \omega \quad \text{with } \alpha_i \in [0,1]$$



Linear SVM path

Optimality condition

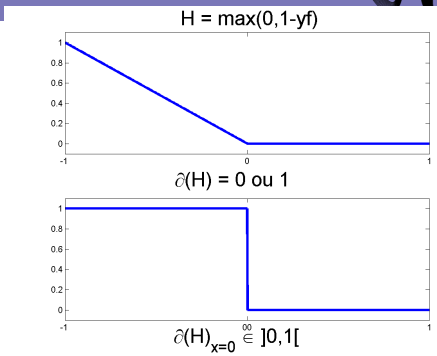
$$\sum_{I_\alpha} \alpha_i y_i x_i^\top + \sum_{I_1} y_i x_i^\top = \lambda\, \omega \quad \text{with } \alpha_i \in [0,1]$$



Path derivation

- Let $\lambda_t \rightarrow$ solution α_i^t, $i \in I_\alpha$, the sets I_α, I_0, I_1
- $\lambda = \lambda_t + \gamma$ such that the sets remain unchanged
- Hence $\forall\, x_j \in I_\alpha,\ f(x_j) = \langle \omega, x_j \rangle = y_j$ (margin points)

Linear SVM path



Optimality condition

$$\sum_{I_\alpha} \alpha_i y_i x_i^\top + \sum_{I_1} y_i x_i^\top = \lambda\, \omega \quad \text{with } \alpha_i \in [0,1]$$

Path derivation

- Let $\lambda_t \rightarrow$ solution α_i^t, $i \in I_\alpha$, the sets I_α, I_0, I_1
- $\lambda = \lambda_t + \gamma$ such that the sets remain unchanged
- Hence $\forall\, x_j \in I_\alpha$, $f(x_j) = \langle \omega, x_j \rangle = y_j$ (margin points)

$$\begin{array}{rcl} \sum_{I_\alpha} \alpha_i^t y_i k(x_i,x_j) + \sum_{I_1} y_i k(x_i,x_j) & = & \lambda_t\, y_j \quad \text{with } k(x_i,x_j) = \langle x_i, x_j \rangle \\ \sum_{I_\alpha} \alpha_i y_i k(x_i,x_j) + \sum_{I_1} y_i k(x_i,x_j) & = & \lambda\, y_j \\ \hline G(\boldsymbol{\alpha} - \boldsymbol{\alpha}^t) & = & (\lambda - \lambda_t)\mathbf{y}_\alpha \quad \text{with} \quad G_{ij} = y_i k(x_i,x_j) \end{array}$$

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^t + (\lambda - \lambda_t)\,\mathbf{w} \qquad \mathbf{w} = G^{-1}\mathbf{y}_\alpha$$

SVM regularization path

Linear variation

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^t + (\lambda - \lambda_t)\,\mathbf{w}$$

The variation holds until the sets change

Event detection

- $x_i \in I_\alpha \rightarrow I_0 \cup I_1$
  α_i goes to 0 or 1
- $x_i \in I_0 \cup I_1 \rightarrow I_\alpha$
  $y_i f(x_i)$ becomes 1

SVM regularization path

Linear variation

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^t + (\lambda - \lambda_t)\,\mathbf{w}$$

The variation holds until the sets change

Event detection

- $x_i \in I_\alpha \rightarrow I_0 \cup I_1$
  α_i goes to 0 or 1
- $x_i \in I_0 \cup I_1 \rightarrow I_\alpha$
  $y_i f(x_i)$ becomes 1

Algorithm

Similar to the algorithm of lasso path

Remark

- Nonlinear case : $\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}}$
- Use the reproducing property $\langle f(\cdot), k(x, \cdot) \rangle$ to derive the previous results

Dealing with the bias term of SVM model

- SVM model : $f(x) = \langle \omega, x \rangle + b$
- Problem : $\min_{\omega,b} \quad \sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \frac{\lambda}{2} \|\omega\|^2$
- Optimality conditions
  - For ω : $\sum_{I_\alpha} \alpha_i y_i x_i^\top + \sum_{I_1} y_i x_i^\top = \lambda\, \omega \quad$ with $\alpha_i \in [0, 1]$
  - For b : $\sum_{I_\alpha} \alpha_i y_i + \sum_{I_1} y_i = 0 \quad$ with $\alpha_i \in [0, 1]$
- Piecewise linear variation

  Let $\alpha_0 = \lambda\, b$. Using the previous analysis, one gets

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \alpha_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^t \\ \alpha_0^t \end{bmatrix} + (\lambda - \lambda_t) \begin{bmatrix} G & \mathbf{1}^\top \\ \mathbf{1} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_\alpha \\ 0 \end{bmatrix}$$

Illustration of SVM regularization path

Nonlinear SVM with gaussian kernel

Parameters α

Decision function

Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Common points between lasso and SVM path

LASSO	SVM
$\beta = 0$	Initialize α
While $I_0 \neq \emptyset$	While $I_1 \neq \emptyset$
Move a variable x_j	Move a point x_j
$I_0 \leftrightarrow I_\beta$	$I_0 \leftrightarrow I_\alpha \leftrightarrow I_1$
compute $\mathbf{w}$	compute $\mathbf{w}$
compute γ	compute γ
$\beta = \beta^t + (\lambda - \lambda_t)\mathbf{w}$	$\alpha = \alpha^t + (\lambda - \lambda_t)\mathbf{w}$

Lesson

Running the path, we select the "good" variables or points and set the other parameters to zero

Why this behavior of sparsity ?

Regularization path and sparsity

Definition : strong homogeneity set (variables)

$$I_0 = \{j \in \{1, ..., d\} \mid \beta_j = 0\}$$

Theorem

Regular if $L(\beta) + \lambda P(\beta)$ differentiable and if $I_0(\mathbf{y}) \neq \emptyset$

$$\forall \varepsilon > 0,\ \exists \mathbf{y}' \in \mathcal{B}(\mathbf{y}, \varepsilon) \text{ such that } I_0(\mathbf{y}') \neq I_0(\mathbf{y})$$

Singular if $L(\beta) + \lambda P(\beta)$ **NON** differentiable and if $I_0(\mathbf{y}) \neq \emptyset$

$$\exists \varepsilon > 0,\ \forall \mathbf{y}' \in \mathcal{B}(\mathbf{y}, \varepsilon) \text{ then } I_0(\mathbf{y}') = I_0(\mathbf{y})$$

singular criteria $\Longrightarrow$ sparsity — Nikolova, 2000

L_1 criteria are singular in 0

singurality provides sparsity

Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation

Extensions of piecewise linear path algorithm

Lasso type

Seminal paper : LAR algorithm [Efron et al. 2004]

- Elastic net (double penalization L_1 and L_2) [Zhou and Hastie, 2005]
- Fused Lasso (L_1 and total variation penalizations) [Tibshrani et al. 2005]
- Grouped Lasso [Yuan and Lin, 2006]
- Least absolute deviation regression (L_1 loss and penalization) [wang et al. 2007]
- Non negative garotte [Yuan and Lin, 2007]
- L_1 penalization in infinite dimension [Rosset et al. 2007]
- Graph data and Lasso [Tsuda, 2007]
- ...

Extensions of piecewise linear path algorithm

SVM type

Seminal paper : SVM path [Efron et al. 2004]

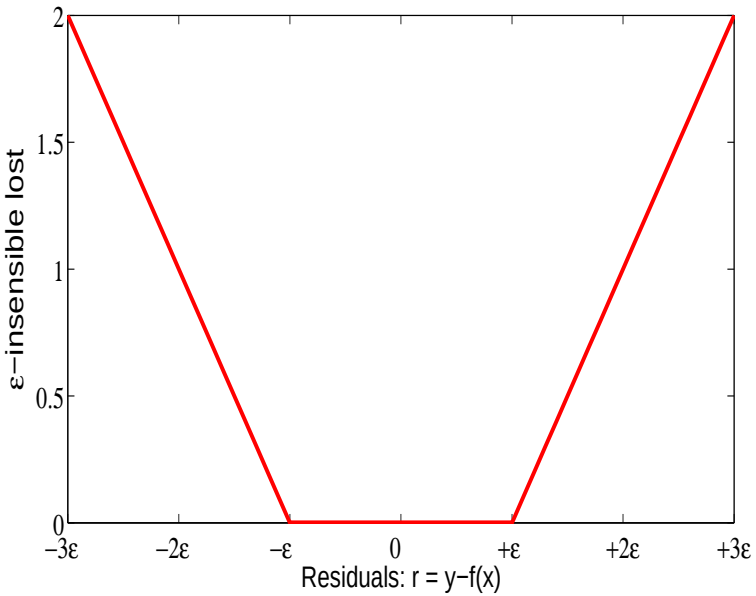
- 1-norm SVM (SVM with L_1 penalty) [Zhou et al. 2003]
- Assymetric cost SVM [Bach et al. 2005]
- Doubly regularized SVM [Wang al. 2006]
- ν-SVM [Loosli et al. 2007]
- SVR [Gunter and Zhu, 2005], [Wang et al. 2006], [Gasso et al., 2007]
- Laplacian Semi-supervised SVM [Wang et al. 2006], [Gasso et al. 2007]
- Oneclass SVM [Rakotomamonjy and Davy 2007]
- Ranking SVM [Zapien et al. 2008]
- ...

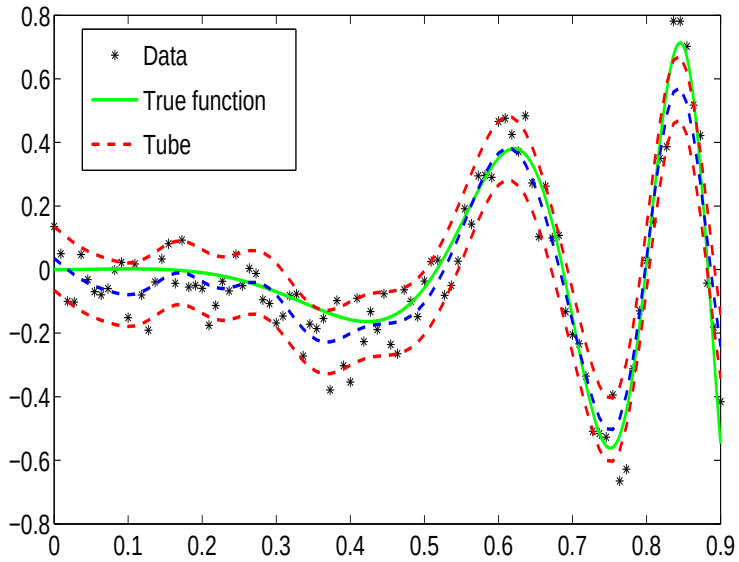
Empirical efficiency evaluation

ν-SVR [Gasso et al. 06]

$$\min_{f,\epsilon} \frac{1}{n}\sum_{i=1}^{n} max(0, |y_i - f(x_i)| - \epsilon) + \nu\epsilon + \frac{\lambda}{2}\|f\|^2 \qquad \text{s.t.} \quad \epsilon \geq 0$$

- Two hyperparameters : ν and ε
- ε insensitive tube with ε : tube width





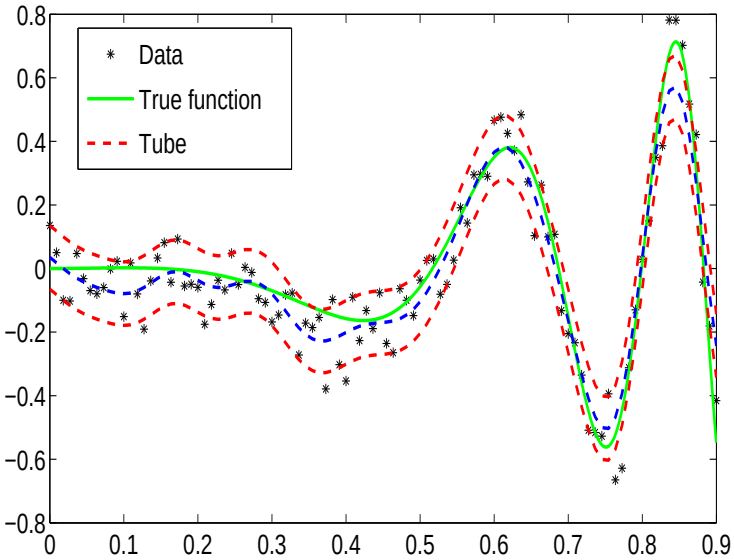
Empirical efficiency evaluation

ν-SVR [Gasso et al. 06]

$$\min_{f,\epsilon} \frac{1}{n}\sum_{i=1}^{n} max(0, |y_i - f(x_i)| - \epsilon) + \nu\epsilon + \frac{\lambda}{2}\|f\|^2 \qquad \text{s.t.} \quad \epsilon \geq 0$$

- Two hyperparameters : ν and ε
- ε insensitive tube with ε : tube width

Toy problem



- Gaussian kernel with bandwidth $\sigma = 0.05$
- Run the λ-path for different values of ν
- Average over 10 trials

Empirical efficiency evaluation

ν-SVR [Gasso et al. 06]

$\min_{f,\epsilon} \frac{1}{n} \sum_{i=1}^{n} max(0, |y_i - f(x_i)| - \epsilon) + \nu\epsilon + \frac{\lambda}{2}\|f\|^2 \qquad \text{s.t.} \quad \epsilon \geq 0$

- Two hyperparameters : ν and ε
- ε insensitive tube with ε : tube width

$N = 1500$ samples - Computational time (sec)

	$\nu = 0.01$	$\nu = 0.5$	$\nu = 0.75$
λ-path	1.70 ± 0.076	1.95 ± 0.03	2 ± 0.031
ν-SVR with warm restart	4.30 ± 0.053	21.8 ± 0.15	21.15 ± 0.12

Computational gain up to 11

Empirical efficiency evaluation

Efficiency of the algorithm : Boston Housing data (UCI repository)

- Multidimensional regression ($x \in \mathcal{R}^{13}$), 506 points
- $N = 406$ samples for training
- Gaussian kernel with different bandwidths σ
- Run the λ-path for different values of ν
- Average over 10 trials (random data selection)

Empirical efficiency evaluation

Efficiency of the algorithm : Boston Housing data (UCI repository)

- Multidimensional regression ($x \in \mathcal{R}^{13}$), 506 points
- $N = 406$ samples for training
- Gaussian kernel with different bandwidths σ
- Run the λ-path for different values of ν
- Average over 10 trials (random data selection)

Bandwidth $\sigma = 1$ - Computational time (sec)

	$\nu = 0.01$	$\nu = 0.5$	$\nu = 0.75$
λ-path	0.95 ± 0.32	1.95 ± 0.35	2.06 ± 1.31
ν-SVR with warm restart	8.6 ± 1.96	13.08 ± 5.17	13.77 ± 5.15

Computational gain up to 9

Empirical efficiency evaluation

Efficiency of the algorithm : Boston Housing data (UCI repository)

- Multidimensional regression ($x \in \mathcal{R}^{13}$), 506 points
- $N = 406$ samples for training
- Gaussian kernel with different bandwidths σ
- Run the λ-path for different values of ν
- Average over 10 trials (random data selection)

$\sigma = 0.1$ - Computational time (sec)

	$\nu = 0.01$	$\nu = 0.5$	$\nu = 0.75$
λ-path	12.31 ± 0.34	12.29 ± 0.44	12.27 ± 0.38
ν-SVR with warm restart	51.44 ± 0.78	51.63 ± 1.24	51.32 ± 0.95

Computational gain up to 4

One-class SVM [Rakotomamonjy et al., 07]

- Levet set estimation

$$\begin{array}{ll} \min_{\beta,\rho,\xi_i} & \frac{\lambda}{2}\|\beta\|^2 + \sum_{i=1}^{n}\xi_i - \lambda\rho \\ st & \xi_i \geq 0, \quad x_i^\top \beta \geq \rho - \xi_i \quad \forall i = 1, \cdots, n \end{array}$$

One-class SVM [Rakotomamonjy et al., 07]

Tab.: Comparing computational time in seconds of alpha seeding and a regularization path approach for computing several level sets

Datasets	# examples	σ	**Alpha Seeding**	**Reg. Path**
credit	653	1	18.1	0.7
		5	21.4	3.8
		10	15.8	4.4
pima	768	1	54.3	0.8
		5	39.8	20.7
		10	25.5	11.2
yeast-cyt	1484	1	42.9	49.42
		5	42.6	51.87
		10	42.5	38.9
spamdata	4601	1	18220	7460
		5	2265	1446
		10	1114	1039

Concluding remarks

Summary

- Linear combination of convex criteria $\longrightarrow$ Pareto frontier $\equiv$ Regularization path
- Efficient computation of the path and sparsity
- Practical for small and medium data set

Extensions

- Large scale data
- Non convex case
- Stopping on the path especially for more than two criteria

References

- F. Bach, D. Heckerman, and E. Horvitz, On the path to an ideal ROC Curve : considering cost asymmetry in learning classifiers, AISTATS, 2005
- B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression, Annals of statistics, vol. 32 (2), pp.407-499, 2004
- G. Gasso, K. Zapien and S. Canu, Computing and stopping the solution paths for ν-SVR, ESANN 2007
- G. Gasso, K. Zapien and S. Canu, Sparsity regularization path for Semi-Supervised SVM, ICMLA 2007
- Gunter and J. Zhu, Computing the solution path for SVM, NIPS 2005.
- G. Loosli, G. Gasso and S. Canu, Regularization paths for ν-SVM and ν-SVR, ISNN 2007
- M. Nikolova, Local strong homogeneity of a regularized estimator, SIAM Journal on Applied Mathematics, vol. 61, no. 2, pp. 633-658, 2000.
- A. Rakotomamonjy, M. Davy, One-class SVM regularization path and comparison with alpha seeding, ESANN 2007
- S. Rosset, Ji Zhu, Piecewise Linear Regularized Solution Paths, Annals of Statistics, 2007
- S. Rosset, G. Swirszcz1 , N. Srebro and Ji Zhu, L_1 Regularization in Infinite Dimensional Feature Spaces, Colt 2007

References

- R. Tibshirani, S. Rosset, Ji Zhu and K. Knight, Sparsity and Smoothness via the Fused Lasso. JRSSB, 2005.
- K. Tsuda, Entire Regularization Paths for Graph Data, ICML 2007
- G. Wang, D.-Y. Yeung, F. H. Lochovsky, Two-Dimensional Solution Path for Support Vector Regression, ICML, 2006
- G. Wang, T. Yeund, and F. H. Lochovsky. Solution path of semi-supervised classification with manifold regularization. ICDM 2006.
- L. Wang, M. Gordon and J. Zhu, Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning, ICDM'06
- L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. Statistica Sinica, 2006
- M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, JRSSB, 2006.
- M. Yuan and Y. Lin. On the non-negative garrotte estimator. Journal of The Royal Statistical Society Series B, 2007.
- K. Zapien, T. Gartner, G. Gasso, and S. Canu, Regularisation Path for Ranking SVM, ESANN 2008
- J. Zhu and T. Hastie, Regularization anf variable selection via the elastic net, JRSSB, 2005.
- J. Zhu, S. Rosset, T. Hastie and R. Tibshirani. 1-Norm Support Vector, NIPS 2003.