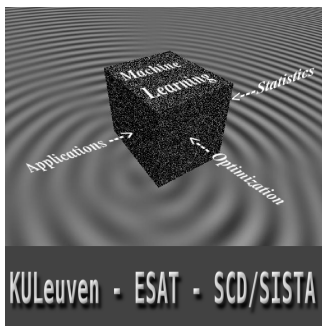


Convex Approaches to Model Selection

K. PELCKMANS, J.A.K. SUYKENS

NIPS workshop on multi-level Inference
December 2006

KULeuven - Department of Electrical Engineering - SCD/SISTA
Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium
Kristiaan.Pelckmans@esat.kuleuven.ac.be



Overview

- Ridge Regression, Smoothing Splines, Regularization networks, LS-SVMs:

$$u? : (\Omega + \gamma I_d) u = v, \quad \Omega \in \mathbb{R}^{d \times d}, \quad \gamma > 0$$

- Solution path when varying γ
- Convex Hull of solution path
- Tuning = learning optimal element in solution path

► Introduction

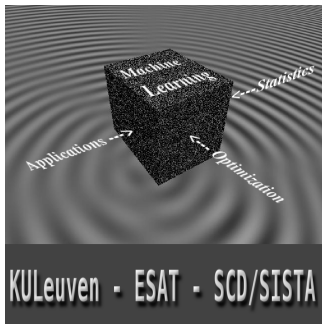
△ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions



Overview

- Ridge Regression, Smoothing Splines, Regularization networks, LS-SVMs:

$$u? : (\Omega + \gamma I_d) u = v, \quad \Omega \in \mathbb{R}^{d \times d}, \quad \gamma > 0$$

- Solution path when varying γ
- Convex Hull of solution path
- Tuning = learning optimal element in solution path

► Introduction

△ Ridge set

△ Tuning ridge

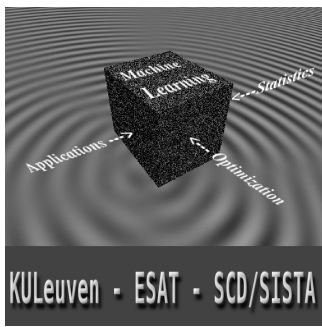
△ w.r.t. CV

△ Example

△ Conclusions

Why convex:

- **(Practice)** Well-developed algorithms
- **(Convexity)** Reproducibility and analyzable
- **(Complexity)** Complexity convex hull
- **(Extensions)** Learning more tuning-parameters
- **(Approach to the global minimum)** projecting on original path



Overview

- Ridge Regression, Smoothing Splines, Regularization networks, LS-SVMs:

$$u? : (\Omega + \gamma I_d) u = v, \quad \Omega \in \mathbb{R}^{d \times d}, \quad \gamma > 0$$

- Solution path when varying γ
- Convex Hull of solution path
- Tuning = learning optimal element in solution path

► Introduction

△ Ridge set

△ Tuning ridge

△ w.r.t. CV

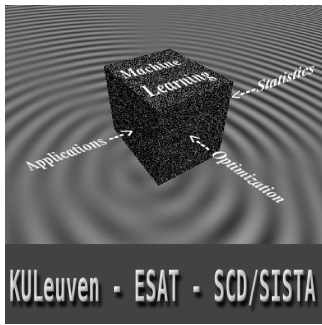
△ Example

△ Conclusions

Why convex:

- **(Practice)** Well-developed algorithms
- **(Convexity)** Reproducibility and analyzable
- **(Complexity)** Complexity convex hull
- **(Extensions)** Learning more tuning-parameters
- **(Approach to the global minimum)** projecting on original path

Pelckmans *et al.*, A Convex Approach to Validation-based Learning of the Regularization



△ Introduction

▶ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions

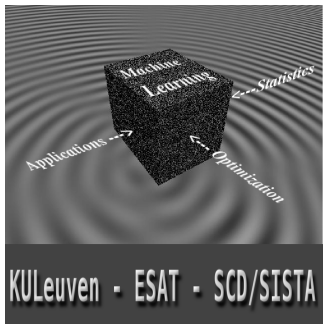
Ridge solution set

- Solution set ('regularization path'):

$$\mathcal{S}(\gamma, u | \Omega, v) = \left\{ u_\gamma \in \mathbb{R}^N \mid \exists 0 < \gamma < +\infty \text{ s.t. } (\Omega + \gamma I_N) u_\gamma = v \right\}$$

- Let $U\Sigma U^T = \Omega$ denote the SVD of the matrix Ω with $UU^T = U^T U = I_N$ and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_N)$ containing all ordered positive eigenvalues such that $\sigma_1 \geq \dots \geq \sigma_N$.
- Rewrite KKT as

$$\left(U\Sigma U^T + \gamma I_n \right) \alpha = Y \Leftrightarrow U_i^T \alpha = \left(\frac{1}{\sigma_i + \gamma} \right) U_i^T Y \quad \forall i$$



Ridge solution set (Ct'd)

△ Introduction

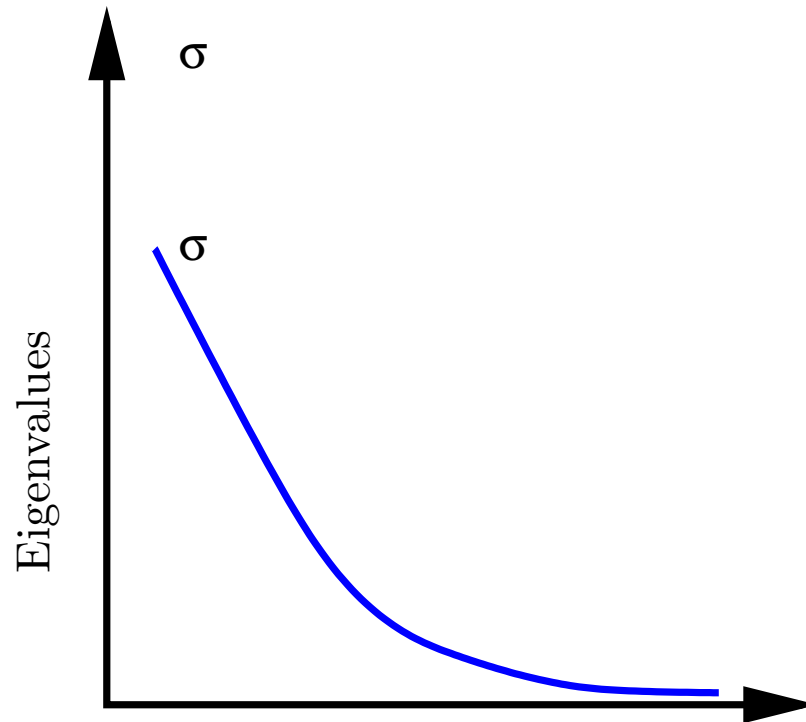
▶ **Ridge set**

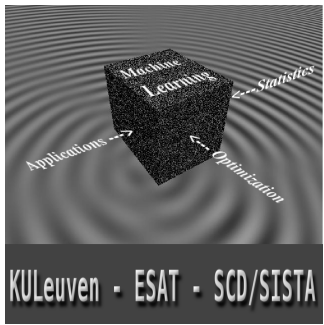
△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions





Ridge solution set ($Ct'd$)

△ Introduction

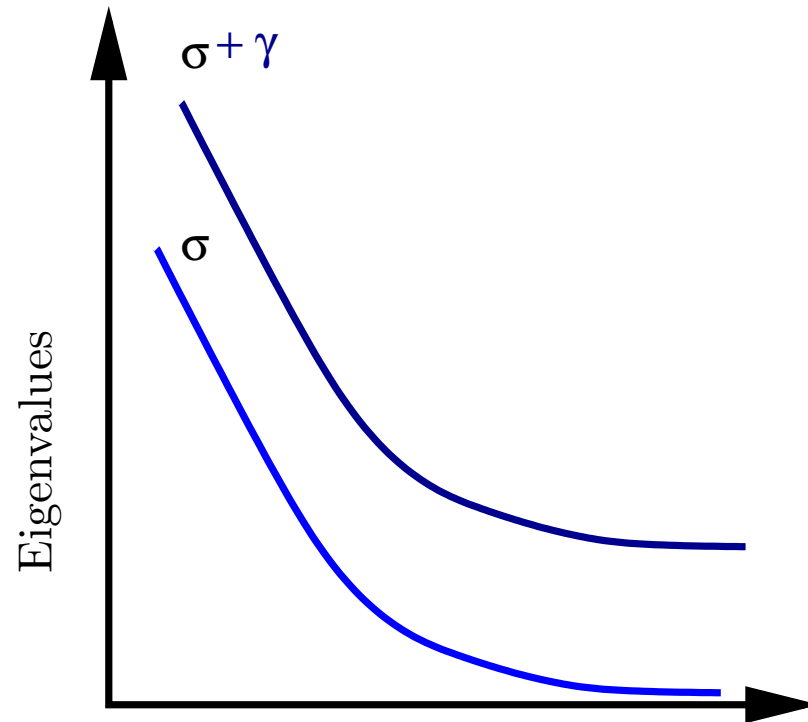
▶ Ridge set

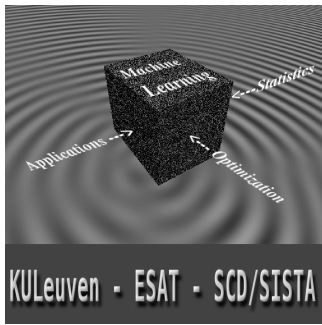
△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions





Ridge solution set ($Ct'd$)

△ Introduction

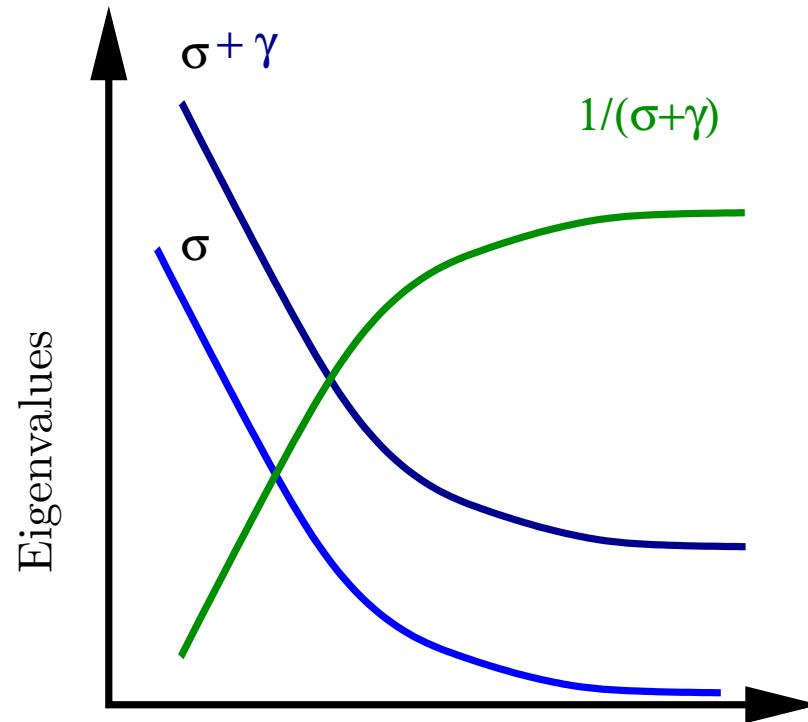
▶ Ridge set

△ Tuning ridge

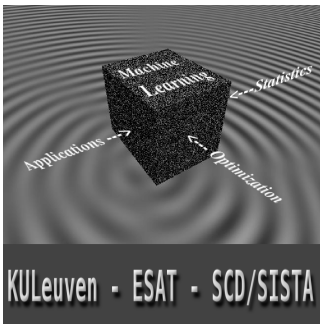
△ w.r.t. CV

△ Example

△ Conclusions



$$\forall \gamma : \frac{1}{\sigma + \gamma} \text{ monotonically increasing in } \sigma$$



Ridge solution set (Ct'd)

△ Introduction

▶ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

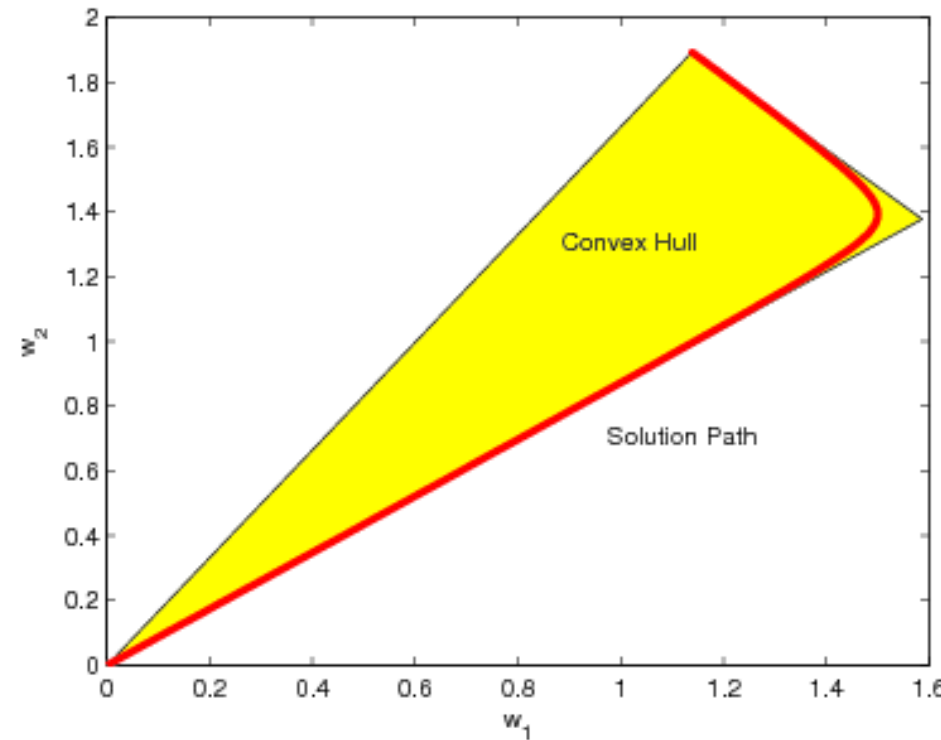
△ Conclusions

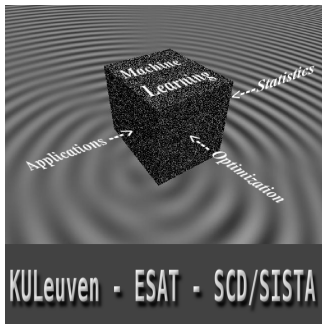
Convex relaxation:

$$\mathcal{S}'(\Lambda, u | \Omega, v)$$

$$= \begin{cases} U_i^T u = \lambda_i U_i^T v & \forall i = \\ 0 < \lambda_i < \frac{1}{\sigma'_i} & \forall i = \\ \left(\frac{\sigma'_k}{\sigma'_i} \right) \lambda_k \leq \lambda_i < \lambda_k & \forall \sigma'_i \\ \lambda_k = \lambda_i & \forall \sigma'_k \end{cases}$$

→ Searching in convex set!





△ Introduction

▶ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions

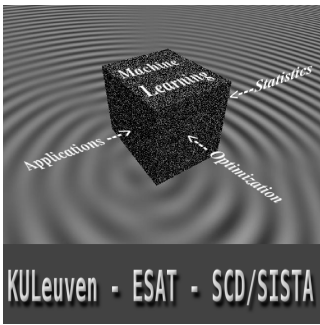
Ridge solution set (Ct'd)

Main result:

Proposition 1. [Maximal distance of relaxation] *the maximal distance from an element in $\mathcal{S}'(\Gamma, u|\Omega, v)$ to its closest counterpart of the non-convex $\mathcal{S}(\gamma, u|\Omega, v)$ can be bounded in terms of the maximum range of the inverse eigenvalue spectrum,*

$$\forall \Gamma \in \mathbb{R}^N \quad \min_{\gamma} \left\| U\Gamma^{-1}U^T v - (\Omega + \hat{\gamma}I_N)^{-1}v \right\|_2 \leq \|v\|_2 \max_{i>k} \left(\left| \frac{1}{\sigma'_i} - \frac{1}{\sigma'_k} \right| \right)$$

Proposition 2. [Smoothness of the Ridge Solution Set] *The solutionset $\mathcal{S}(\gamma, u|\Omega, v)$ is Lipschitz smooth when $\sigma_N > 0$.*



Ridge solution set (Ct'd)

Corollary 1. [Modified Ridge Regression yielding a Convex Solution Path]

The convex relaxation constitutes the solution path for the modified ridge regression problem

$$\hat{w} = \arg \min_w \mathcal{J}_\Gamma(w) = \sum_{i=1}^n (w^T x_i - y_i)^2 + \frac{1}{2} w^T (U\Gamma U^T) w$$

where $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_D)$ and γ_d satisfies the constraint $\gamma_d = \frac{1}{\lambda_d} - \sigma_d$ for all $d = 1, \dots, D$, and the following inequalities hold:

$$\begin{cases} \gamma_d > 0 & \forall d = 1, \dots, D \\ \left(\frac{\sigma_g}{\sigma_d}\right) (\sigma_d + \gamma_d) \geq (\sigma_g + \gamma_g) > (\sigma_d + \gamma_d) & \forall \sigma_g > \sigma_d \\ \gamma_d = \gamma_g & \forall \sigma_d = \sigma_g \end{cases}$$

△ Introduction

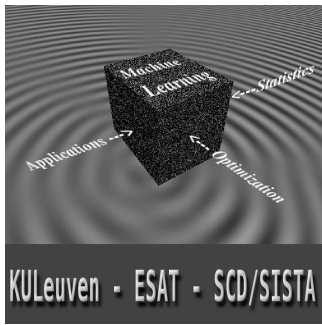
▶ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

△ Conclusions



Tuning γ in Ridge Regression

Data $\{(x_i, y_i)\}_{i=1}^n$ and $\{(x_j^v, y_j^v)\}_{j=1}^m$ iid from F_{XY}

Model $f(x) = w^T x$, training problem:

$$\hat{w} = \arg \min_w \mathcal{J}_\gamma(w) = \sum_{i=1}^n (w^T x_i - y_i)^2 + \frac{\gamma}{2} w^T w$$

Normal equations:

$$\text{KKT}(w|\gamma, \mathcal{D}) : (X^T X + \gamma I_D) w = X^T Y$$

Tuning the regularization constant:

$$(\hat{w}, \hat{\gamma}) = \arg \min_{w, \gamma > 0} \sum_{j=1}^{n_v} \ell(w^T x_j^v - y_j^v) \quad \text{s.t.} \quad \text{KKT}(w|\gamma, \mathcal{D}) \leftrightarrow \mathcal{S}(\gamma, w|\mathcal{D})$$

Convex relaxation

$$(\hat{w}, \hat{\Lambda}) = \arg \min_{w, \Lambda} \sum_{j=1}^{n_v} \ell(w^T x_j^v - y_j^v) \quad \text{s.t.} \quad \mathcal{S}'(\Lambda, w|\mathcal{D})$$

Pelckmans *et al.*, Additive Regularization Trade-off: Fusion of Training and Validation

△ Introduction

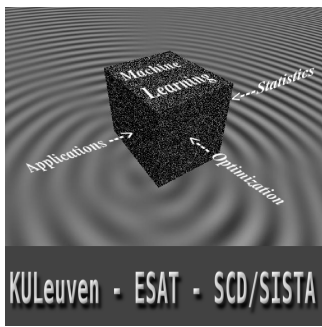
△ Ridge set

▶ **Tuning ridge**

△ w.r.t. CV

△ Example

△ Conclusions



Tuning $\gamma > 0$ in Least Squares Support vector Machines

- Predictive model $f(x) = w^T \varphi(x)$ with $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^D$
- Training problem:

$$\min_{w,e} \mathcal{J}_\gamma(w, e) = \sum_{i=1}^n e_i^2 + \frac{\gamma}{2} w^T w \text{ s.t. } w^T \varphi(x_i) + e_i = y_i, \forall i$$

- Normal equations:

$$\text{KKT}(\alpha|\gamma, \mathcal{D}) : (K + \gamma I_n) \alpha = Y$$

- Optimal Prediction model $\hat{f}(x) = \hat{w}^T \varphi(x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$
- Tuning the regularization constant:

$$(\hat{w}, \hat{\gamma}) = \arg \min_{w, \gamma > 0} \sum_{j=1}^{n_v} \ell \left(\sum_{i=1}^n \hat{\alpha}_i K(x_i, x_j^v) - y_j^v \right) \text{ s.t. } \mathcal{S}(\gamma, \alpha|\mathcal{D})$$

△ Introduction

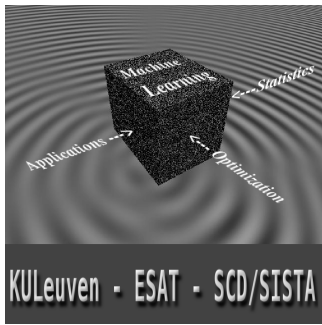
△ Ridge set

► **Tuning ridge**

△ w.r.t. CV

△ Example

△ Conclusions



Tuning the ridge w.r.t. 10-fold CV

L-fold \rightarrow each fold has its own KKT condition:

$$\begin{cases} \text{KKT} (w_{(1)} | \gamma, \mathcal{D}_{(1)}) \\ \text{KKT} (w_{(2)} | \gamma, \mathcal{D}_{(2)}) \\ \vdots \\ \text{KKT} (w_{(10)} | \gamma, \mathcal{D}_{(10)}) \end{cases}$$

...but γ coupled over folds!

Moreover, for each fold exists a separate validation set, thus

$$(\hat{w}_{(l)}, \hat{\gamma}) = \arg \min_{w_{(l)}, \gamma > 0} \sum_{l=1}^L \frac{1}{(n - n_{(l)})} \sum_{(x_i, y_i) \in \mathcal{D}_{(l)}^v} \ell (w_{(l)}^T x_i - y_i)$$

$$\text{s.t. } \text{KKT} (w_{(l)} | \gamma, \mathcal{D}_{(l)}), \forall l = 1, \dots, L$$

△ Introduction

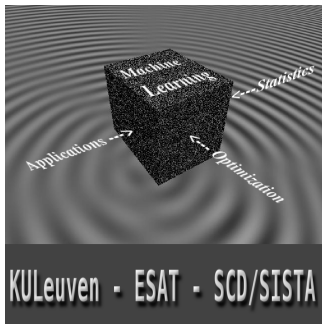
△ Ridge set

△ Tuning ridge

▶ **w.r.t. CV**

△ Example

△ Conclusions



Tuning the ridge w.r.t. CV

Proposition 3. [Coupling over different folds] Let $U_l \Sigma_l U_l^T$ be the SVD of Ω_l for all $l = 1, \dots, 10$. (...) Then the following coupled relaxation is proposed:

$$\mathcal{S}' \left(\Lambda^L, w_{(l)} \mid \mathcal{D}_{(1)}, \dots, \mathcal{D}_{(L)} \right) = \begin{cases} U_i^{(l)T} w_{(l)} = \lambda_k U_i^{(l)T} X^{(l)T} Y^{(l)} & \forall k \leftrightarrow (l), i \\ 0 < \lambda_k < \frac{1}{\sigma_k'} & \forall k = 1, \dots \\ \left(\frac{\sigma_k'}{\sigma_l'} \right) \lambda_k < \lambda_l < \lambda_k & \forall \sigma_l' > \sigma_k' \\ \lambda_k = \lambda_l & \forall \sigma_k' = \sigma_l' \end{cases}$$

Thus the convex relaxation to tuning the ridge with respect to an L -fold CV criterion yields to solving

$$\min_{w_{(l)}, \Lambda} \sum_{l=1}^L \left(\frac{1}{n - n_{(l)}} \right) \sum_{(x_i, y_i) \in \mathcal{D}_{(l)}^v} \ell \left(w_{(l)}^T x_i^v - y_i^v \right)$$

$$\text{s.t. } \mathcal{S}' \left(\Lambda^L, w_{(l)} \mid \mathcal{D}_{(1)}, \dots, \mathcal{D}_{(L)} \right).$$

△ Introduction

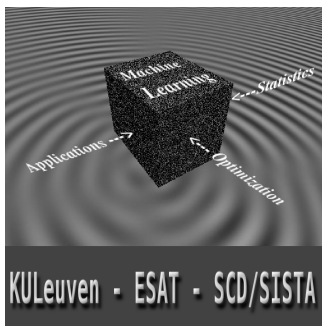
△ Ridge set

△ Tuning ridge

▶ **w.r.t. CV**

△ Example

△ Conclusions



Examples

△ Introduction

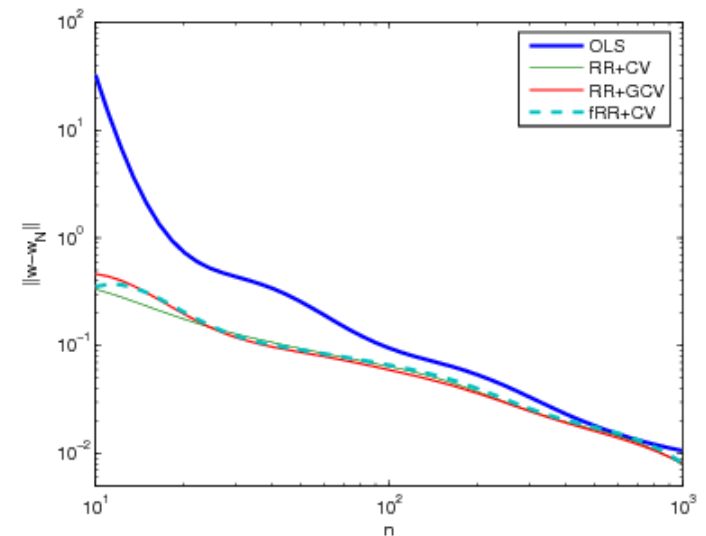
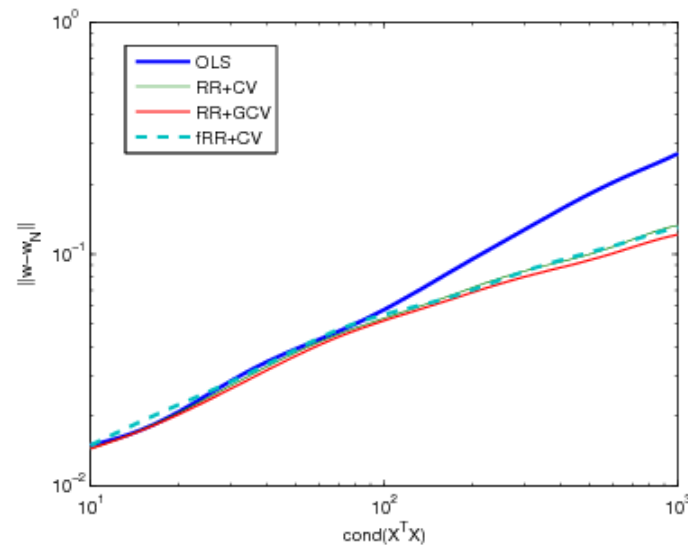
△ Ridge set

△ Tuning ridge

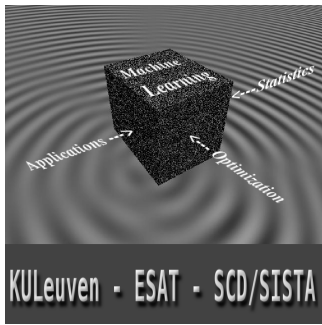
△ w.r.t. CV

▶ **Example**

△ Conclusions



Results of a comparison between OLS and RR with $D = 10$, tuned by CV (steepest descent), by GCV (using steepest descent) and the proposed method fusing training and tuning the ridge together in one convex optimization algorithm. Panel (a) shows the evolution when ranging the condition number with $n = 50$ fixed. Panel (b) displays the evolution of the performance when the number of examples ranges and $\Gamma(X^T X) = 1e^3$ is fixed. In both cases the proposed convex relaxation is performing similar as steepest descent based counterparts, while it significantly outperforms OLS in the case of low n or a high enough condition number $\Gamma(X^T X)$.



△ Introduction

△ Ridge set

△ Tuning ridge

△ w.r.t. CV

△ Example

► **Conclusions**

Conclusions

Message:

- Hyper-parameter tuning as a convex optimization problem
- setting the stage ...
- Convex hull of solution path
- Bounded maximal distance between solution path and relaxation if $\sigma_N > 0$ or $\gamma > 0$
- Efficient tuning w.r.t. validation and CV

Outlooks:

- Input selection for additive models
- Comparison with gradient descent/Bayesian inference
- Convex relaxation of solution path SVM