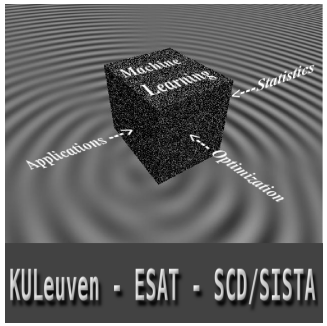
Convex Approaches to Model Selection

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

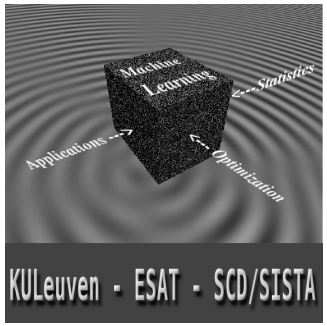


Overview

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

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

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



Overview

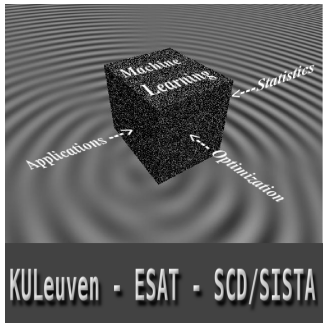
- Ridge Regression, Smoothing Splines, Regularization networks, LS-SVMs:

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

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

Why convex:

- **(Practice)** Well-developed algorithms
- **(Convexity)** Reproducability 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:

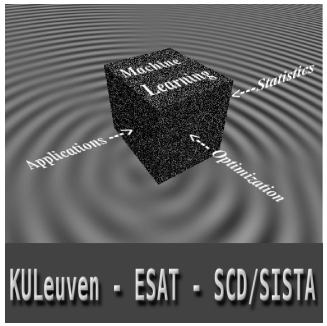
$$u? : \quad (\Omega + \gamma I_d)\ u = v, \quad \Omega \in \mathbb{R}^{d \times d},\ \gamma > 0$$

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

Why convex:

- **(Practice)** Well-developed algorithms
- **(Convexity)** Reproducability 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 Regularizati



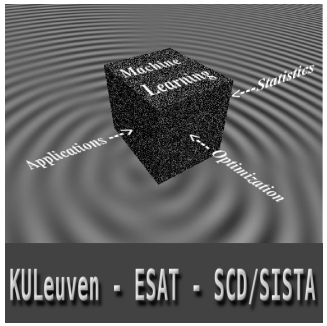
Ridge solution set

- Solution set ('regularization path'):

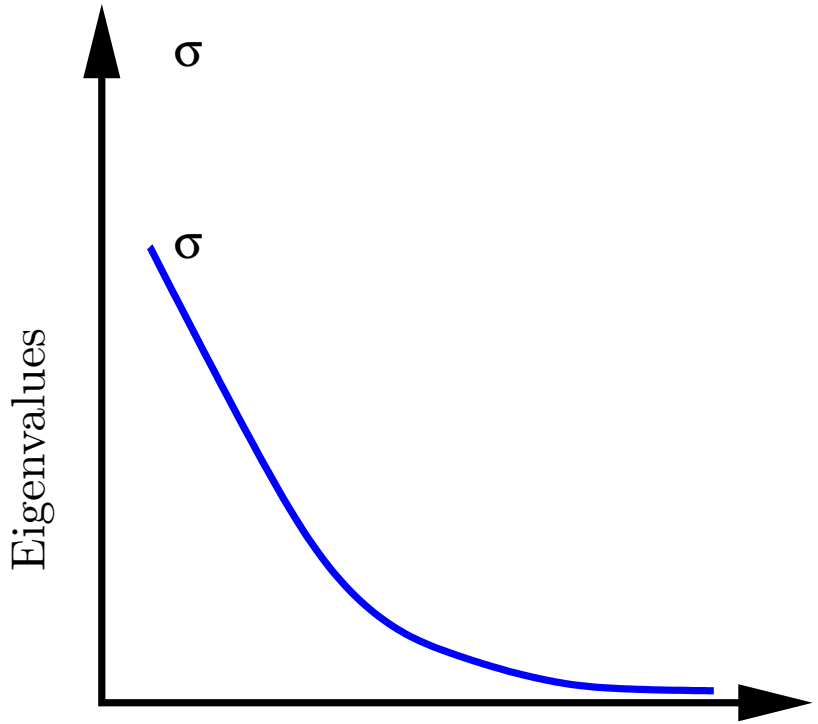
$$\mathcal{S}(\gamma, u|\Omega, v) = \left\{ u_\gamma \in \mathbb{R}^N \;\middle|\; \exists\, 0 < \gamma < +\infty \quad \text{s.t.} \quad (\Omega + \gamma I_N)\, u_\gamma = v \right\}$$

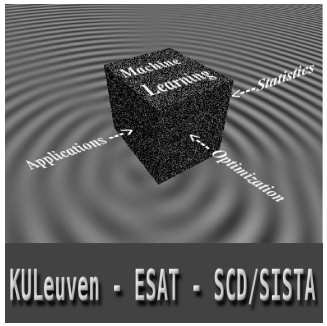
- Let $U\Sigma U^T = \Omega$ denote the SVD of the matrix Ω with $UU^T = U^T U = I_N$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$ containing all ordered positive eigenvalues such that $\sigma_1 \geq \cdots \geq \sigma_N$.
- Rewrite KKT as

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

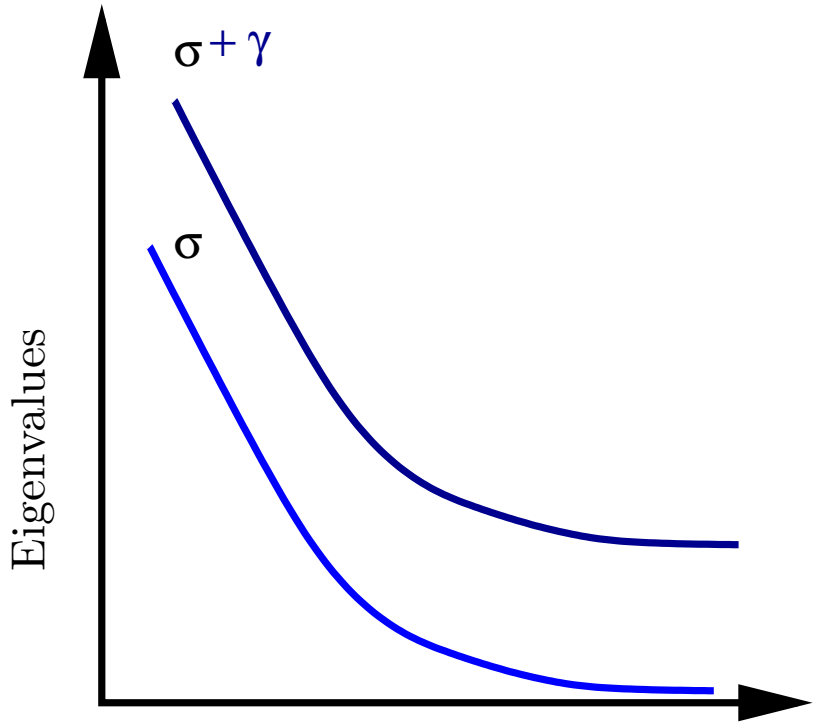


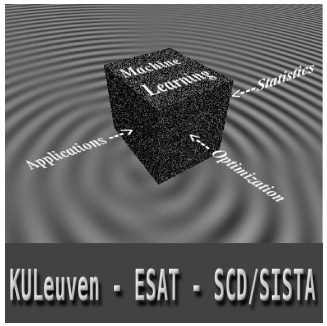
Ridge solution set (Ct'd)



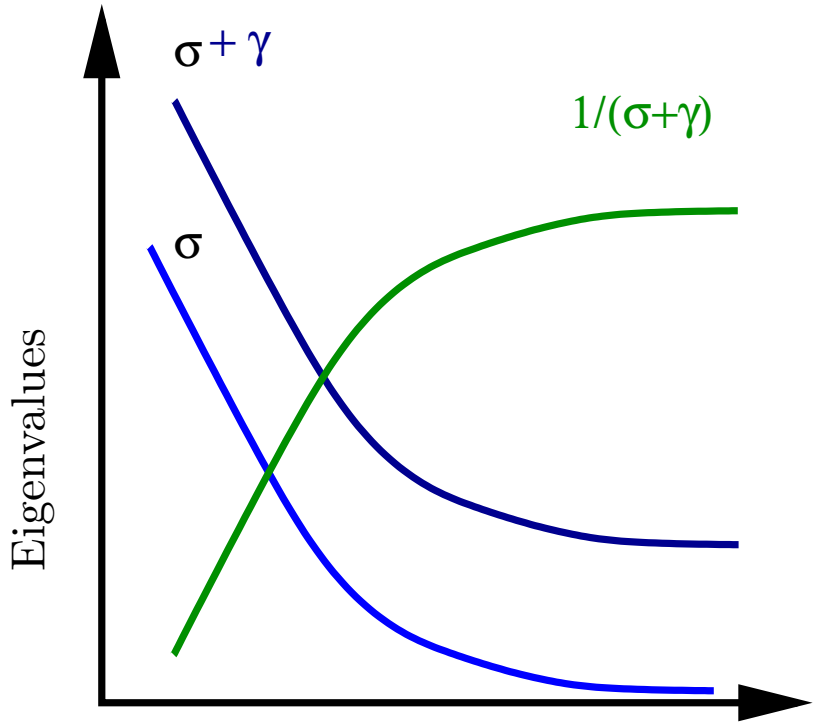


Ridge solution set (Ct'd)

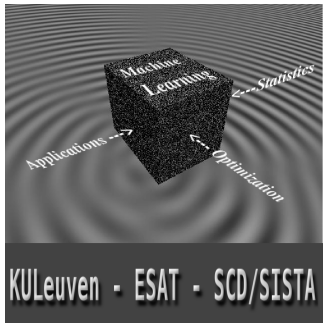




Ridge solution set (Ct'd)



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

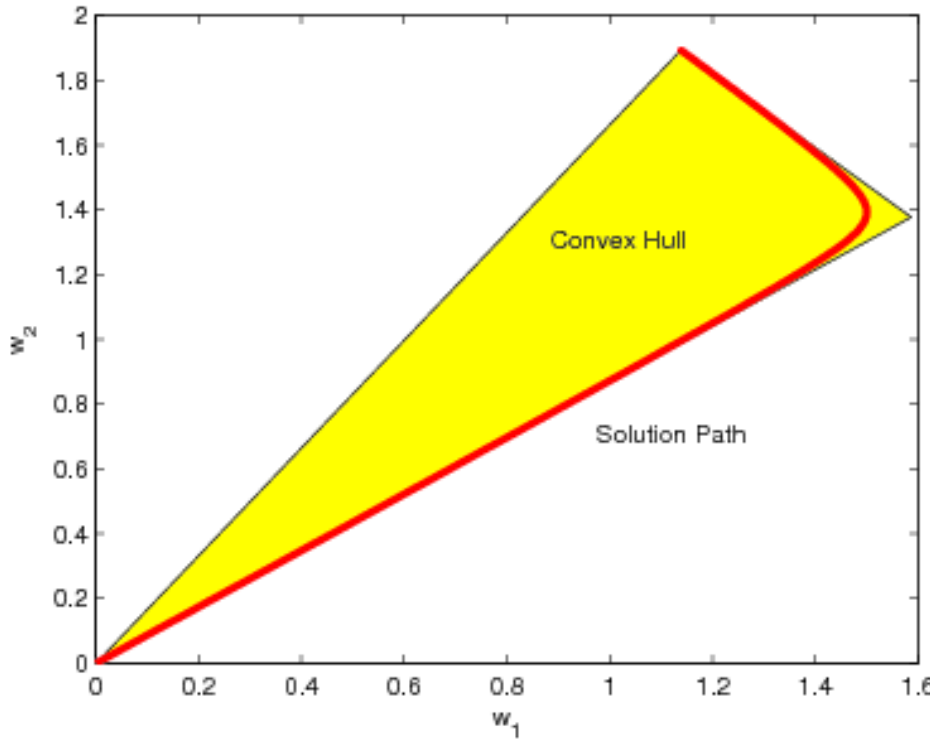


Ridge solution set (Ct'd)

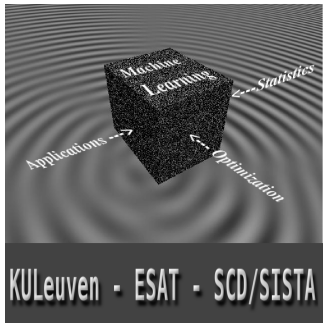
Convex relaxation:

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

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



$\longrightarrow$ Searching in convex set!



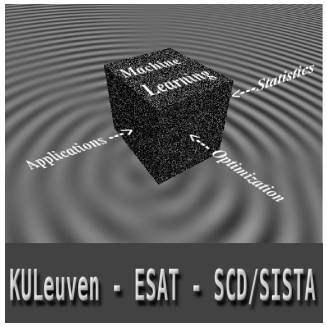
Ridge solution set (Ct'd)

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* $\sigma_N > 0$.



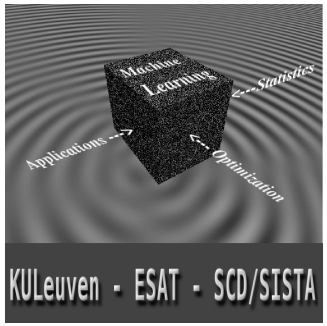
Ridge solution set (Ct'd)

Corollary 1. [Modified Ridge Regression yielding a Convex Solution Path] *The convex relaxation constitutes the solution path for the modified ridge regression problem*

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

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

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



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:

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

Normal equations:

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

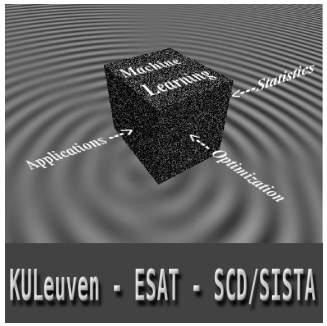
Tuning the regularization constant:

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

Convex relaxation

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

Pelckmans *et al.*, Additive Regularization Trade-off: Fusion of Trainining and Validati



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

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

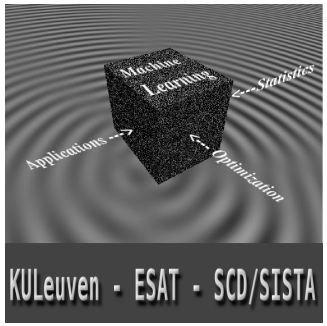
$$\min_{w,e} \mathcal{J}_\gamma(w, e) = \sum_{i=1}^{n} e_i^2 + \frac{\gamma}{2} w^T w \text{ s.t. } \quad w^T \varphi(x_i) + e_i = y_i, \ \forall i$$

- 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:

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



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

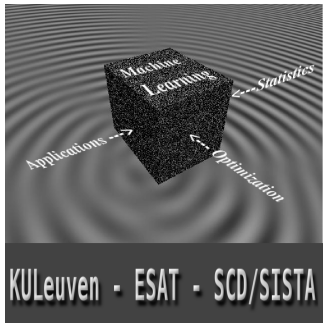
L-fold $\longrightarrow$ each fold has its own KKT condition:

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

...but γ coupled over folds!
Moreover, for each fold exists a separate validation set, thus

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

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



Tuning the ridge w.r.t. CV

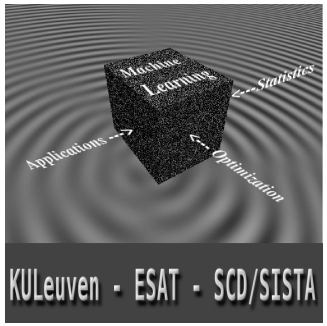
Proposition 3. [Coupling over different folds] *Let $U_l \Sigma_l U_l^T$ be the SVD of Ω_l for all $l = 1, \ldots, 10$. (...) Then the following coupled relaxation is proposed:*

$$\mathcal{S}'\left(\Lambda^L, w_{(l)} \middle|\, \mathcal{D}_{(1)}, \ldots, \mathcal{D}_{(L)}\right) = \begin{cases} {U_i^{(l)}}^T w_{(l)} = \lambda_k {U_i^{(l)}}^T {X^{(l)}}^T Y^{(l)} & \forall k \leftrightarrow (l), i \\ 0 < \lambda_k < \frac{1}{\sigma'_k} & \forall k = 1, \ldots \\ \left(\frac{\sigma'_k}{\sigma'_l}\right) \lambda_k < \lambda_l < \lambda_k & \forall \sigma'_l > \sigma'_k \\ \lambda_k = \lambda_l & \forall \sigma'_k = \sigma'_l \end{cases}$$

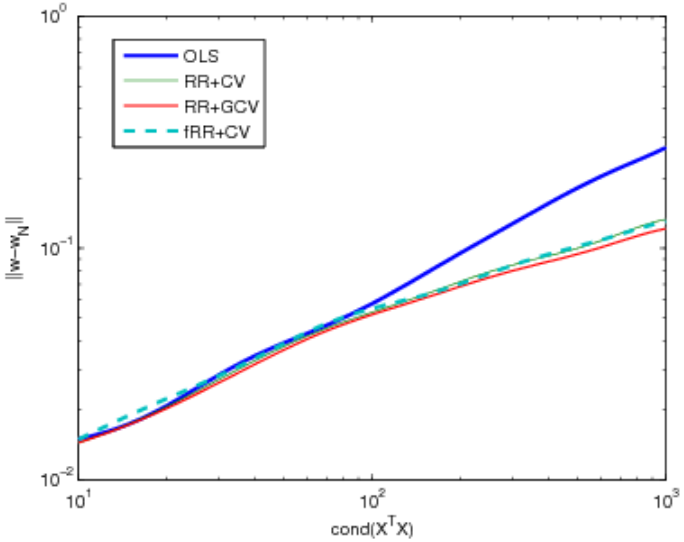
Thus the convex relaxation to tuning the ridge with respect to an L-fold CV criterion yields to solving

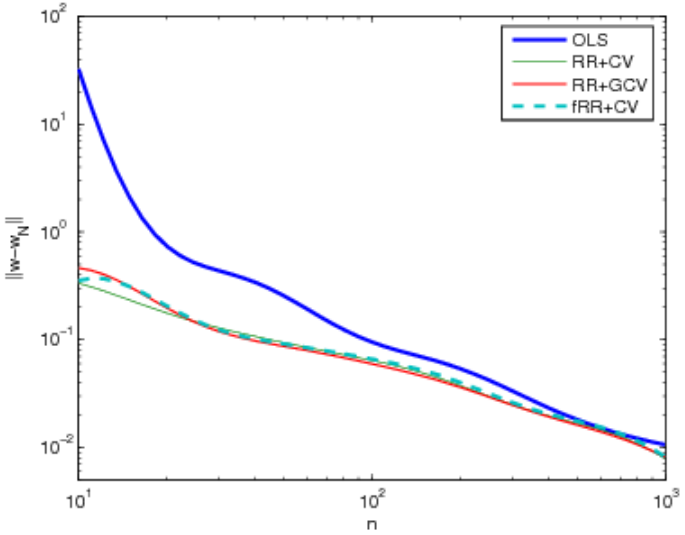
$$\min_{w_{(l)}, \Lambda} \sum_{l=1}^{L} \left(\frac{1}{n - n_{(l)}}\right) \sum_{(x_i, y_i) \in \mathcal{D}^v_{(l)}} \ell\left(w_{(l)}^T x_i^v - y_i^v\right)$$

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

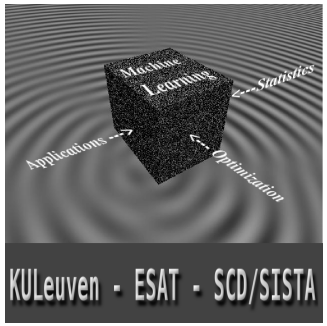


Examples





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 $\Gamma(X^T X)$.*



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