regularized regression: we want to minimize the test error (generalization error). Usually, the test error consists of ~~contr~~ several contributions, for example bias and variance.

Observation: the biggest error source dominates the total ("a chain is only as strong as it's weakest link"). It makes no sense to make all errors except for one small - waste of resources.

All error sources should contribute about equally.

In OLS: bias = 0, variance can be high if features are (nearly) redundant (i.e. linearly dependent).

=> allow some bias if this reduces the variance a lot.

<u>regularization</u>: add a new term to the loss fct. that reduces overfitting

original Loss = $\underbrace{\text{data term}}$, $\boxed{\text{new Loss} = \underbrace{\text{data term}} + \tau \cdot \underbrace{\text{regularization term}}}$

how well does the model with parameters β fit the training set?

prior knowledge: how should models for this problem class typically look like? which parameters β are plausible?

<u>ridge regression</u>: data term $\hat{=}$ squared loss $\|Y - X\beta\|^2$

regularization term $\hat{=}$ L2 regularizer: $\|\beta\|_2^2 = \beta^T\beta$

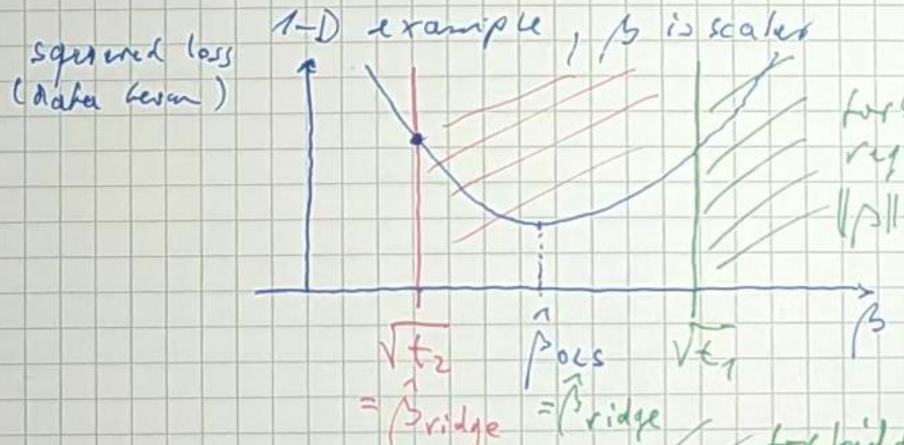
interpretation in term of prior knowledge: - data term $\hat{=}$ negative log-likelihood of the TS => regularization is the neg. log likelihood of $\exp(-\tau\,\beta^T\beta)$

=> prior belief: good parameters are Gaussian distributed with mean 0 and variance $\sigma^2 = \frac{1}{2\tau}$

graphical illustration of ridge regr. $\hat{\beta} = \underset{\beta}{\arg\min} \, \|Y - X\beta\|^2 + \tau \|\beta\|_2^2$

equivalent to $\hat{\beta} = \underset{\beta}{\arg\min} \, \|Y - X\beta\|^2$ s.t. $\|\beta\|_2^2 \le t$

1-D example, β is scalar

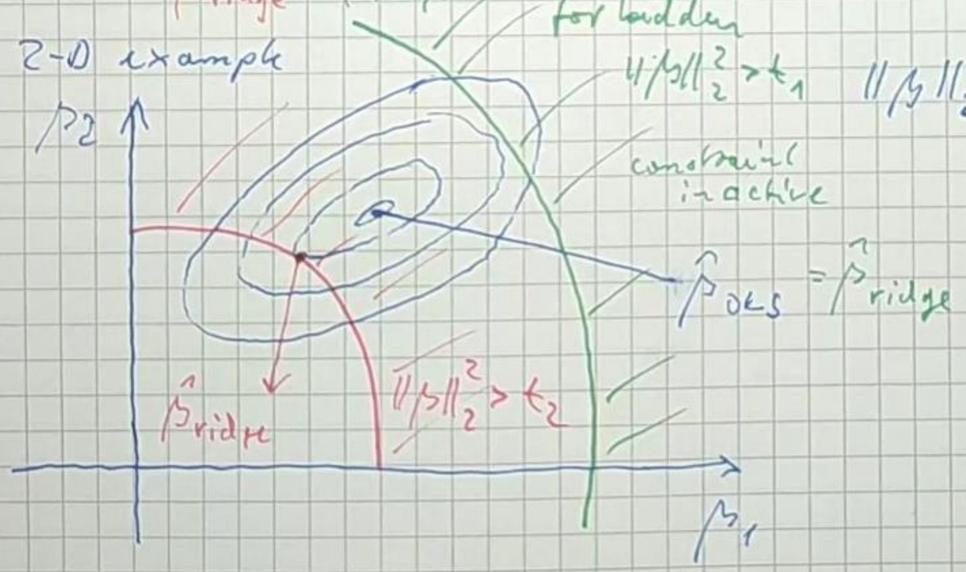


has no effect here, because $\|\beta_{OLS}\|^2 < t_1$, "constraint inactive"

$t_2 < t_1$: $\|\beta_{OLS}\|^2 > t_2 \Rightarrow$ constraint active

$\hat{\beta}_{ridge} = \sqrt{t_2}$

2-D example



$\|\beta\|_2^2 \le t$ defines a circle with radius $\sqrt{t}$ and centered ~~at~~ $\beta = 0$

other possibilities for regularization: feature selection, L_1 regularization

<u>feature selection</u>: idea: when feature are redundant, only use the most relevant non-redundant subset = "active set" $A \subseteq \{1, \dots, D\}$

objective: $$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 \quad \text{s.t.} \quad \underbrace{|A|}_{\text{number of active features}} \leq t$$

(only $\beta_j \neq 0$ when $j \in A$, otherwise $\underbrace{\beta_j = 0}_{\text{inactive coefficients}}$)

"0-norm": $\|\beta\|_0$ = # non-zero coefficients in β = $|A|$

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 + \tau \|\beta_0\|$$

advantage: $\hat{\beta}$ is the exact OLS solution for the active features (= unbiased)

disadvantage: L_0 regularization is NP-hard, $\Rightarrow$ the active set for some t can be very different than for $t-1$ $\Rightarrow$ in general, finding the globally optimal active set for some t requires <u>exhaustive search</u>

in practice, things are usually not so bad: optimal active sets for $t-1$ and t are usually very similar $\Rightarrow$ efficient approximation alg.

<u>orthogonal matching pursuit</u> (OMP): (0) start with $A = \emptyset$

(1) for $m = 1, \dots, t$: add one feature per iteration!

(a) compute ~~squared~~ residuals for current guess: $R_i = \left(Y_i - X_i \overset{\substack{\text{OLS solution for } A^{(m-1)} \\ \downarrow}}{\hat{\beta}^{(m-1)}}\right)$

(b) find the best inactive feature: $\hat{j} = \arg\max_{j \notin A^{(m-1)}} |X_j^T R|$ and add: $A^{(m)} = A^{(m-1)} \cup \{\hat{j}\}$

L1 regularization   LASSO regression ("least absolute shrinkage and selection operator")

in between L_2 regularization $\|\beta\|_2^2 \le t$ and L_0 regul. $\|\beta\|_0 \le t$

it shrinks coefficients towards 0 like L_2 reg., but not as much

selects features like L_0 reg., but is not NP-hard, but convex

objective: $\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2$ s.t. $\|\beta\|_1 \le t$

$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 + \tau \|\beta\|_1$ $\|\beta\|_1 = \sum_{j=1}^{D} |\beta_j|$

convex => unique solution, easy to find, various alg., for example

LARS algorithm: similar to OMP, but after each iteration, check if $A^{(m)}$ has become redundant after adding the new feature (in practice, this occurs every once in a while, but not so very often)

if yes => remove the least important feature from $A^{(m)}$ before adding a new one

in 1-D, LASSO and ridge regression are identical, because $\|\beta\|_2 = \|\beta\|_1$

=> same diagram