Goal : To improve model performance by reducing overfitting
To improve model interpretability by reducing coefficients of irrelevant predictors to zero (or close to zero).

How: Change the cost/loss function.

- <u>Ridge regression</u>

$$\text{Loss} = RSS + \underset{\text{hyperparameter}}{\lambda} \overbrace{\sum_{j=1}^{p} \beta_j^2}^{\text{reduces overfitting}}$$

- Increasing λ decreases model flexibility $\Rightarrow$ reduces overfitting.
(increases bias reduces variance)

- All features need to be standardized because in vanilla linear regression, changing scale of a predictor $(c \times x_j)$ led to automatic adjustment of the coefficient $(\frac{1}{c} \times \beta_j)$ so that $\beta_j x_j$ was constant.
But, for ridge, the loss function also has β_j^2 term so $\beta_j \cdot x_j$ won't be constant and will lead to biased estimate of β_j if x_j is not standardized.

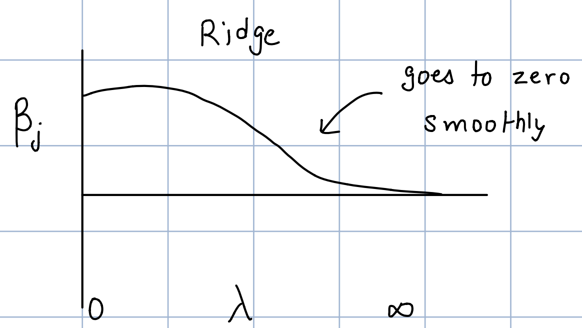
- Does not shrink β_js to 0.

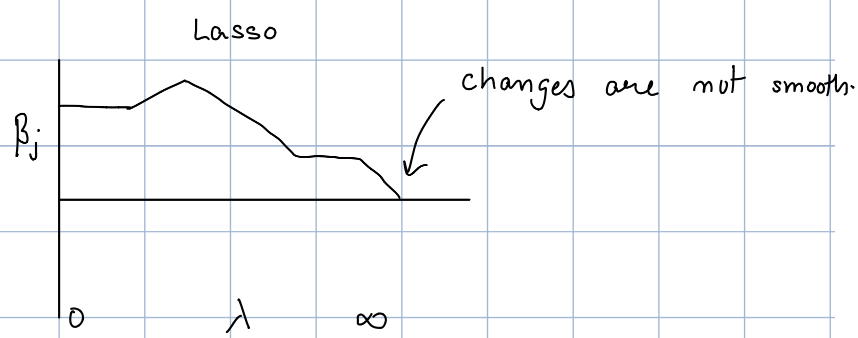
- ## Lasso regression

$$\text{Loss} = RSS + \underset{\downarrow \text{ hyperparameter}}{\lambda} \overbrace{\sum_{j=1}^{p} |\beta_j|}^{\text{reduces overfitting AND shrinks some } \beta_j\text{s to } 0}$$

- Increasing λ decreases model flexibility $\Longrightarrow$ reduces overfitting. (increases bias reduces variance)
- Variable selection: shrinks coefficients of some predictors to exactly zero, therefore removing these from the model.

- ## Comparing Ridge and Lasso





Lasso is better when a lot of variables do not relate to Y.
Check CV (val) error to see which one is better.

Another formulation of Ridge and Lasso:

Ridge: $\underset{\beta}{\text{minimize}}$ RSS subject to $\sum \beta_j^2 \leq s$

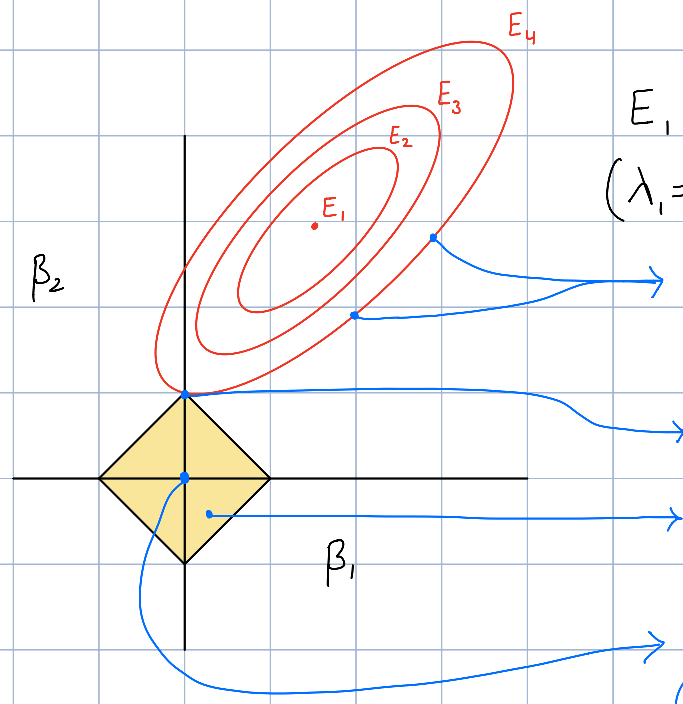
Lasso: $\underset{\beta}{\text{minimize}}$ RSS subject to $\sum |\beta_j| \leq s$

Both equations have multiple solutions for β that fulfill the constraints.

E_1 = corresponds to βs that give min RSS (error) on training set

$$E_1 < E_2 < E_3 < E_4 < \dots$$
$$(\lambda_1 = 0 < \lambda_2 < \lambda_3 < \lambda_4 < \dots)$$

Lasso:
$(p = 2)$



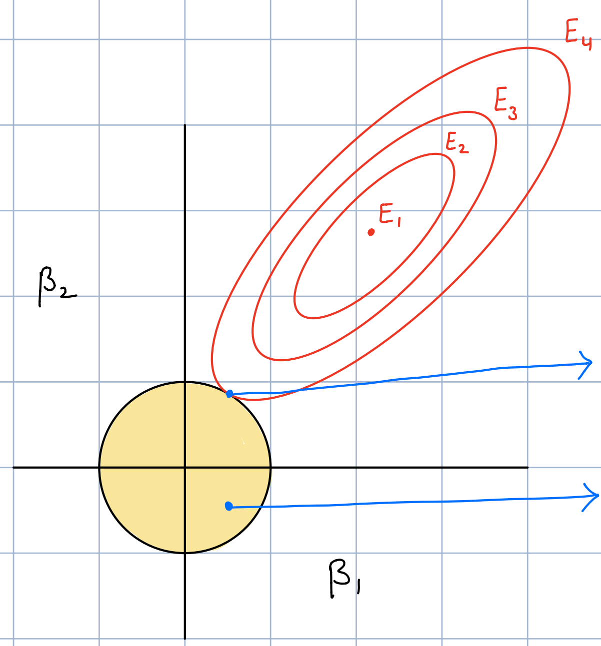
Multiple solutions with same error (and λ)

solution that shrinks β_1 to 0.

Solution in the region that satisfies $|\beta_1| + |\beta_2| \le s$ (L_1 norm region)

Max regularization
$(\lambda = \infty \,;\, |\beta_1| + |\beta_2| \le 0)$

Ridge:
$(p = 2)$



$\beta_1 \neq \beta_2 \neq 0$ (but β_1 is close to zero)

L_2 norm region
Satisfies $\beta_1^2 + \beta_2^2 \le s$

$$L_p\text{-norm} = \left(\sum_i |x_i|^p \right)^{1/p}$$

L_1-norm (Lasso): $\sum_i |\beta_i| \le s$

L_2-norm (Ridge): $\sum_i \beta_i^2 \le r$ where $r = s^2$

Note: Choose λ by trying out various values and picking the one that gives lowest CV error on validation set.