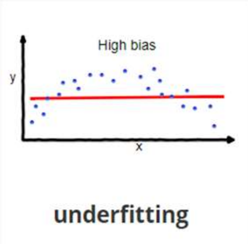
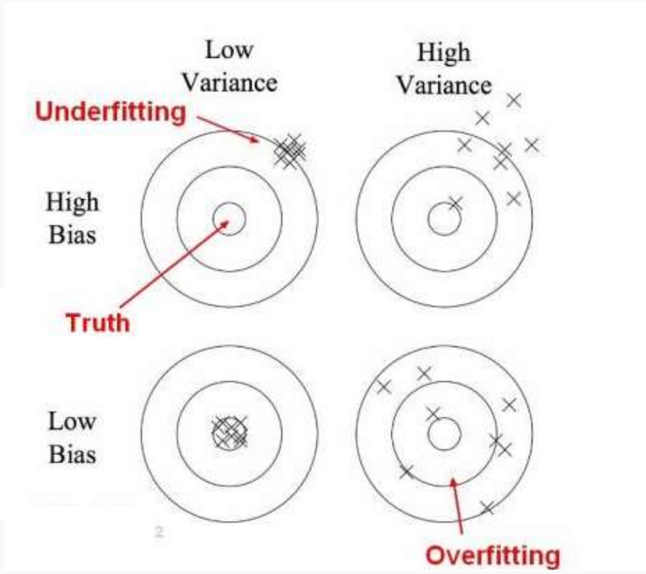
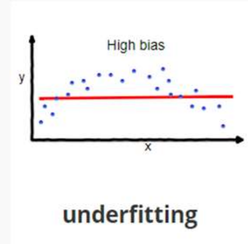
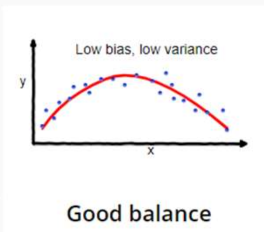
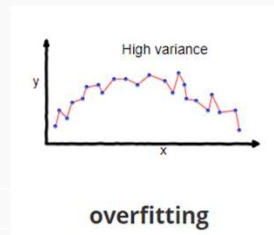
# Motivation




**Regularization**

1

## Regularization: An Overview

The idea of regularization revolves around modifying the loss function L; in particular, we add a regularization term that penalizes some specified properties of the model parameters

$$L_{reg}(\beta) = L(\beta) + \lambda R(\beta),$$

where $\lambda$ is a scalar that gives the weight (or importance) of the regularization term.

Fitting the model using the modified loss function $L_{reg}$ would result in model parameters with desirable properties (specified by $R$).

2

## LASSO Regression (Least absolute shrinkage and selection operator)

Since we wish to discourage extreme values in model parameter, we need to choose a regularization term that penalizes parameter magnitudes. For our loss function, we will again use MSE.

Together our regularized loss function is:

$$L_{LASSO}(\beta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda \sum_{j=1}^{J} |\beta_j|.$$

Note that $\sum_{j=1}^{J} |\beta_j|$ is the $l_1$ norm of the vector $\beta$

$$\sum_{j=1}^{J} |\beta_j| = \|\boldsymbol{\beta}\|_1$$

3

## Ridge Regression

Alternatively, we can choose a regularization term that penalizes the squares of the parameter magnitudes. Then, our regularized loss function is:

$$L_{Ridge}(\beta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda \sum_{j=1}^{J} \beta_j^2.$$

Note that $\sum_{j=1}^{J} |\beta_j|^2$ is the square of the $l_2$ norm of the vector $\beta$

$$\sum_{j=1}^{J} \beta_j^2 = \|\boldsymbol{\beta}\|_2^2$$

4

## Choosing $\lambda$

In both ridge and LASSO regression, we see that the larger our choice of the **regularization parameter** $\lambda$, the more heavily we penalize large values in $\beta$,

- If $\lambda$ is close to zero, we recover the MSE, i.e. ridge and LASSO regression is just ordinary regression.
- If $\lambda$ is sufficiently large, the MSE term in the regularized loss function will be insignificant and the regularization term will force $\beta_{\text{ridge}}$ and $\beta_{\text{LASSO}}$ to be close to zero.

To avoid ad-hoc choices, we should select $\lambda$ using cross-validation.

5

## Regularization Parameter with a Validation Set

The solution of the Ridge/Lasso regression involves three steps:

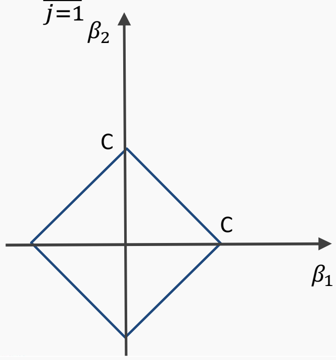
- Select $\lambda$
- Find the minimum of the ridge/Lasso regression loss function (using the formula for ridge) and record the *MSE* **on the validation set**.
- Find the $\lambda$ that gives the smallest *MSE*

6

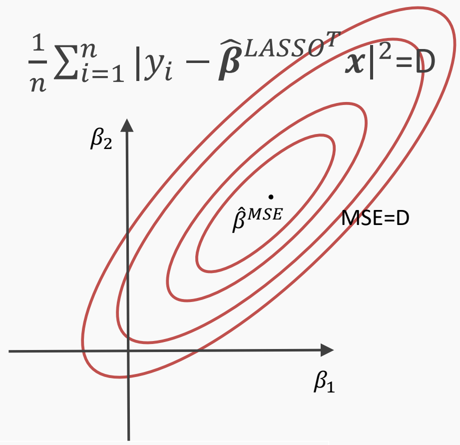
## The Geometry of Regularization (LASSO)

$$L_{LASSO}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \boldsymbol{\beta}^T \boldsymbol{x}\right|^2 + \lambda \sum_{j=1}^{J} |\beta_j| \qquad \widehat{\boldsymbol{\beta}}^{LASSO} = \operatorname{argmin} L_{LASSO}(\boldsymbol{\beta})$$

$$\lambda \sum_{j=1}^{J} |\,\hat{\beta}_j^{LASSO}| = C$$




$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\boldsymbol{\beta}}^{LASSO^T} \boldsymbol{x}|^2 = D$$

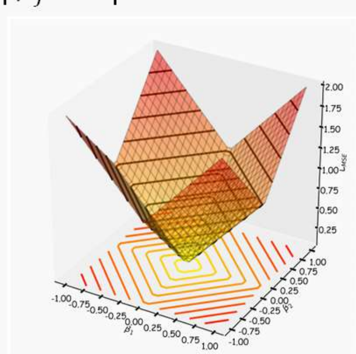



7

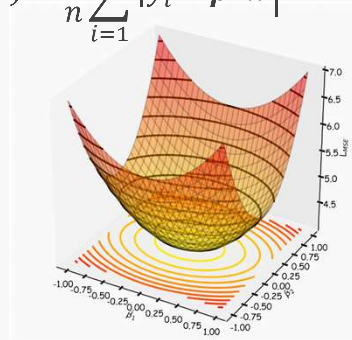
## The Geometry of Regularization (LASSO)

$$L_{LASSO}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \boldsymbol{\beta}^T \boldsymbol{x}\right|^2 + \lambda \sum_{j=1}^{J} |\beta_j| \qquad \widehat{\boldsymbol{\beta}}^{LASSO} = \operatorname{argmin} L_{LASSO}(\boldsymbol{\beta})$$

$$L_1 = \lambda \sum_{j=1}^{J} |\,\hat{\beta}_j^{LASSO}|$$

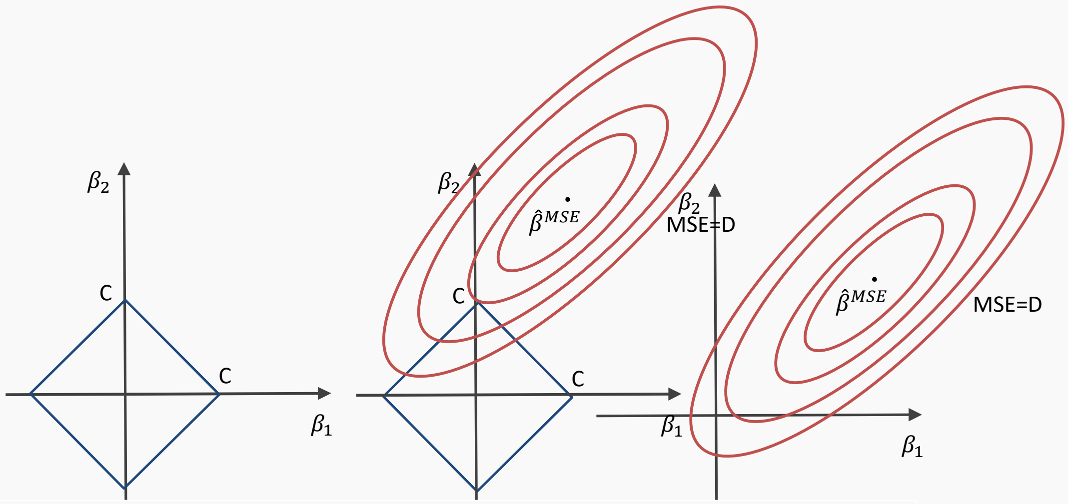


$$L_{MSE}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \boldsymbol{\beta}^T \boldsymbol{x}\right|^2$$



8

## The Geometry of Regularization (LASSO)

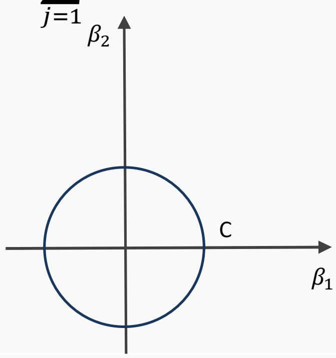



9

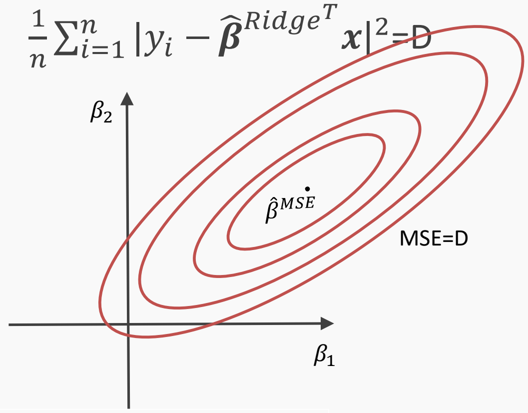
## The Geometry of Regularization (Ridge)

$$L_{Ridge}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \boldsymbol{\beta}^T \boldsymbol{x}\right|^2 + \lambda \sum_{j=1}^{J} (\beta_j)^2 \qquad \widehat{\boldsymbol{\beta}}^{Ridge} = \operatorname{argmin} L_{Ridge}(\boldsymbol{\beta})$$

$$\lambda \sum_{j=1}^{J} \left|\hat{\beta}_j^{Ridge}\right|^2 = C$$

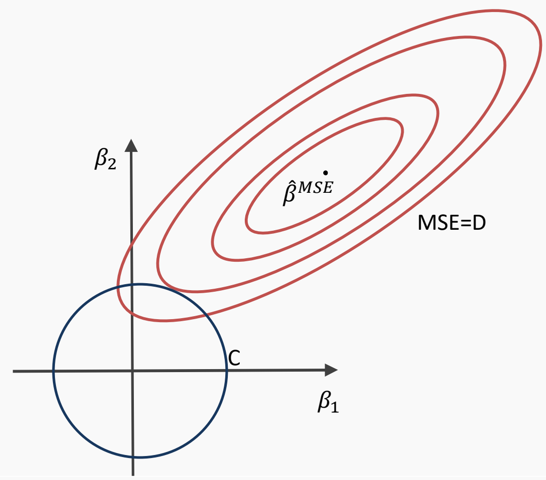



$\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\boldsymbol{\beta}}^{Ridge^T} \boldsymbol{x}|^2$=D



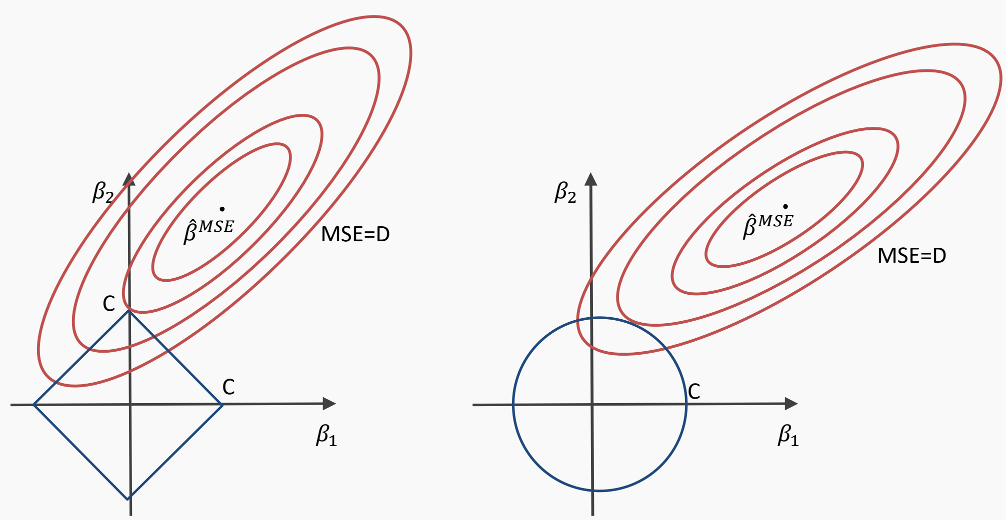

10

## The Geometry of Regularization (Ridge)




11

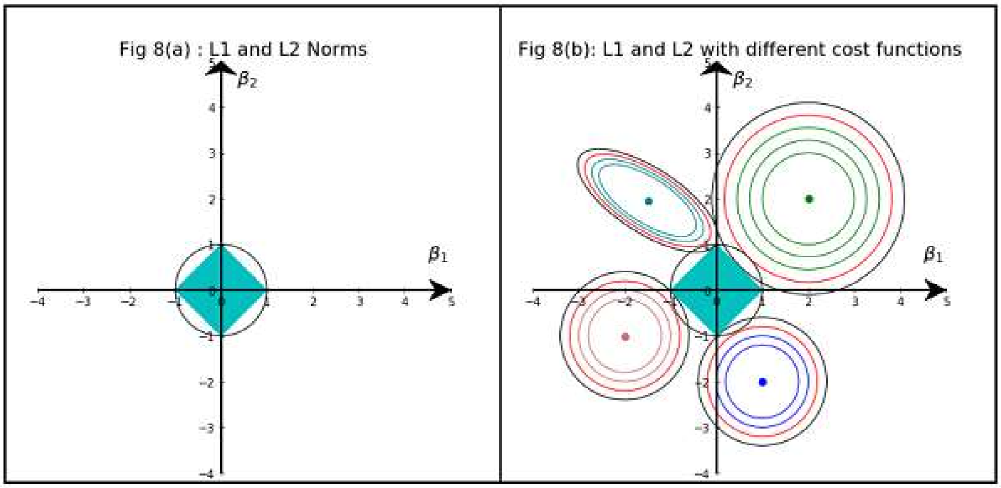
## The Geometry of Regularization




12

## The Geometry of Regularization

**LASSO: Coefficient estimates that are exactly equal to zero**




13

## Comparison of Lasso and Ridge Regression

| Feature | Lasso Regression | Ridge Regression |
|---|---|---|
| **Penalty Term** | **L1 norm** (sum of absolute values of coefficients) | **L2 norm** (sum of squared coefficients) |
| **Impact on Coefficients** | Shrinks some coefficients to exactly **zero**, effectively selecting key features. | Shrinks coefficients **towards zero** but does not eliminate any features. |
| **Feature Selection** | Yes, Lasso performs **automatic feature selection**, resulting in a sparse model. | No, Ridge retains all features but reduces their impact by shrinking coefficients. |
| **Best Suited For** | Situations where some features are irrelevant, and reducing dimensionality is beneficial. | Cases with multicollinearity where all features contribute but need controlled influence. |
| **Computational Complexity** | Can be computationally demanding due to the non-differentiability of the L1 norm. | Computationally efficient as the L2 norm is differentiable. |
| **Bias-Variance Tradeoff** | Introduces **higher bias** but reduces variance, making the model simpler. | Results in **lower bias** but maintains some variance, preserving more information. |

14