L1 and L2 regularization for feature selection

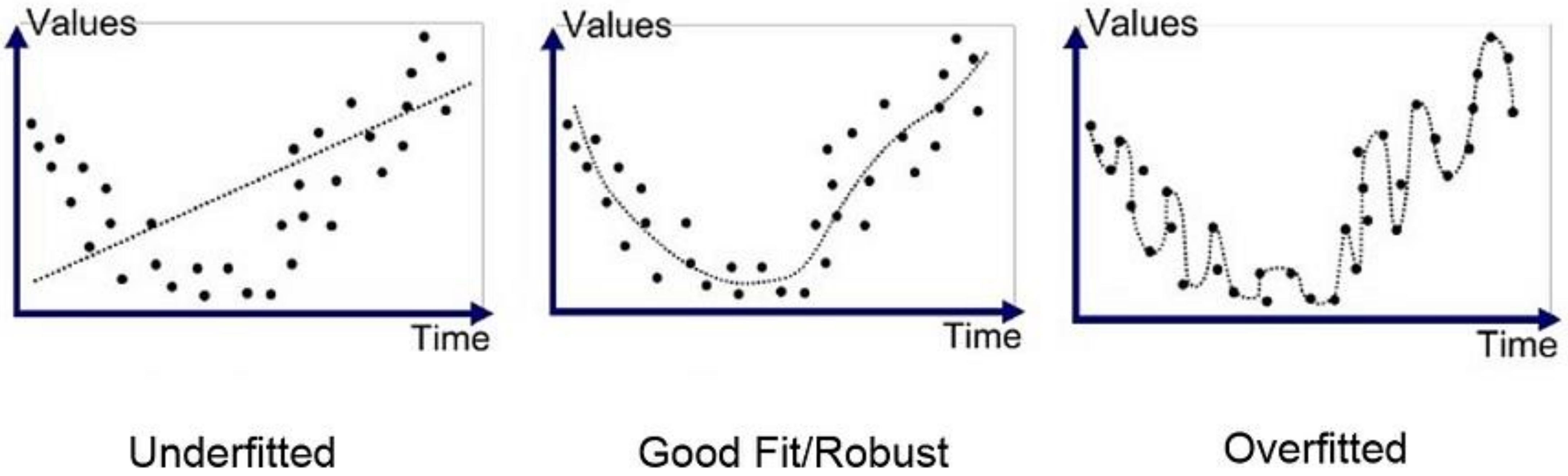
Week 05 Advanced Topic 2

NTNU

Overview

- Motivation
- Theoretical part
- Examples
- Discussion
- Conclusions

Overfitting



- Generalization Problems
- High Variance
- Captures Noise

https://medium.com/swlh/machine-learning-how-to-prevent-overfitting-fdf759cc00a9

Common Solutions to Overfitting

- **Train with more data**
  - Use more, accurate training data → more accurate and well-fitted model
  - **Challenges:**
    - Expensive and impractical.
    - Data quality matters (noisy or irrelevant data can worsen performance)
- **Early Stopping**
  - For iterative models, stops training when validation error stops improving
  - **Challenges:**
    - Requires separate validation dataset
    - "Stopping point" not always clear

https://medium.com/swlh/machine-learning-how-to-prevent-overfitting-fdf759cc00a9

L1/L2 Regularization

Prevents overfitting by controlling model complexity and reducing large coefficients (model weights).

How?

- During training, the **loss function** measures how well the model fits the data, and the goal is to minimize this.
- **Regularization adds a penalty to the loss function** to discourage the model from becoming too complex.

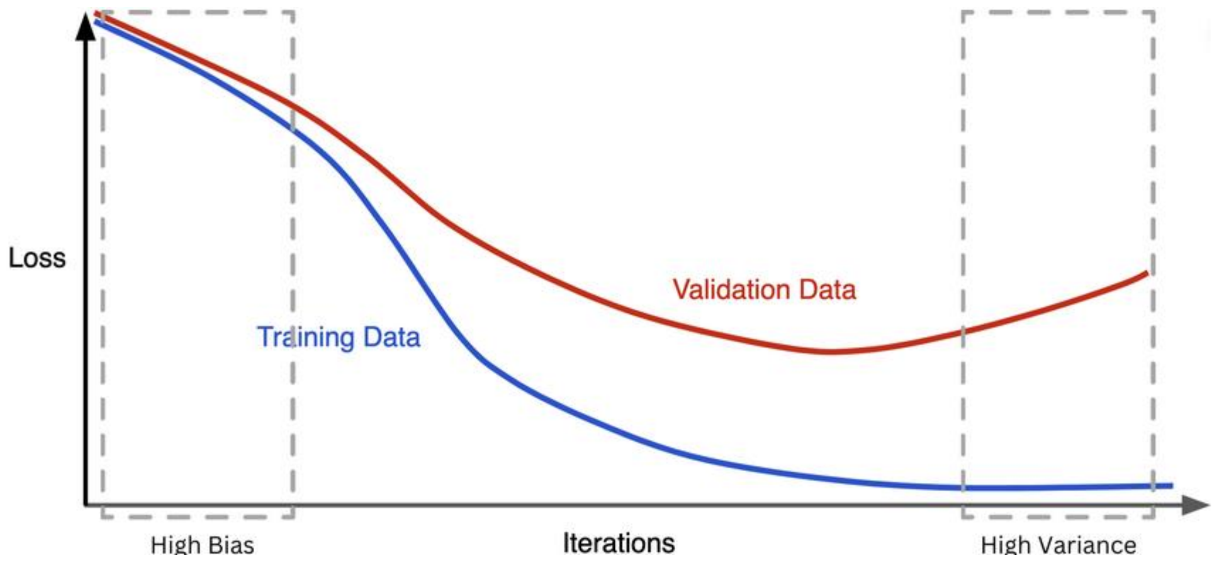
$$\text{Regularization} = \text{Loss Function} + \text{Penalty}$$

- This **penalty pushes less important weights closer to zero**, simplifying the model and preventing overfitting.

L1/L2 - What does the penalty represent?

- **Recap: The bias-variance tradeoff**
  - **High bias:** Simpler model
  - **High variance:** Higher chance of overfitting

- **What is penalty**
  - The penalty **adds slight bias** to the loss function.
  - This **reduces variance** and simplifies the model, preventing overfitting.



L1 Regularization

L1 Regularization

Theoretical part

- Selection Operator (Lasso).
- Adds a penalty term to the cost function.
- Feature selection.

Lasso Regression Cost Function = Loss Function + $\lambda \sum_{\{j=1\}}^{m} |\beta_j|$

Controls the strength of regularization

Model parameters

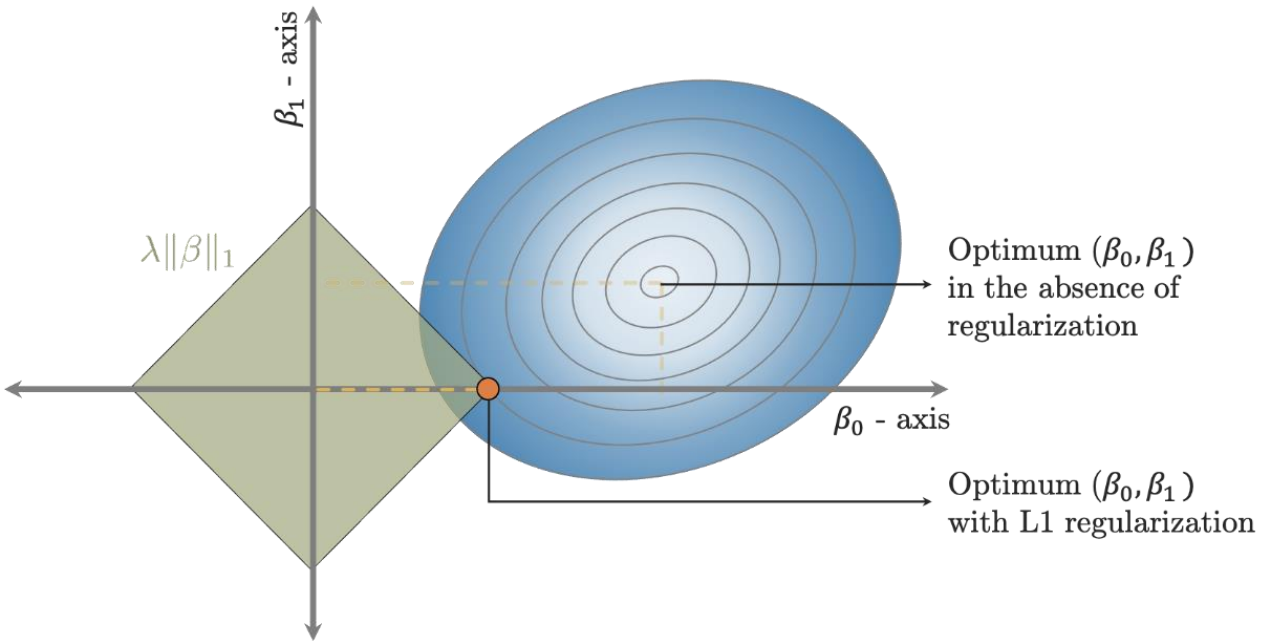


Fig.2: L1 regularization example.

NTNU

L1 Illustration

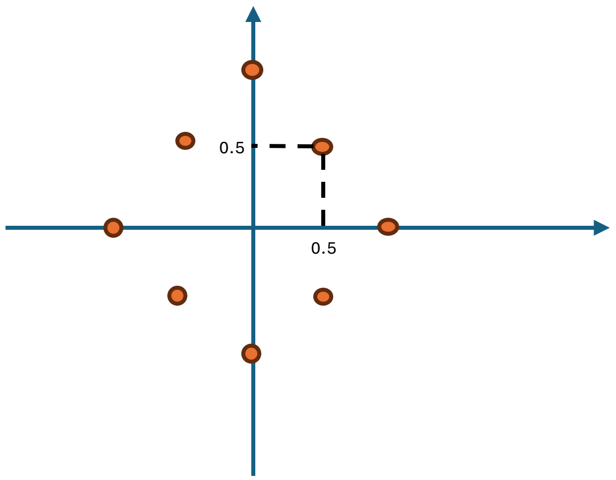
Theoretical part

Let's make a unit "ball":

$$\sum_{\{j=1\}}^{m} |\beta_j| = 1$$

$$|\beta_1| + |\beta_2| = 1$$

$$|1| + |0| = 1$$
$$|0.5| + |0.5| = 1$$
$$\ldots$$



L1 Regularization

Theoretical part

- Selection Operator (Lasso).
- Adds a penalty term to the cost function.
- Feature selection.

Lasso Regression Cost Function = Loss Function + $\lambda \sum_{\{j=1\}}^{m} |\beta_j|$

Controls the strength of regularization

Model parameters

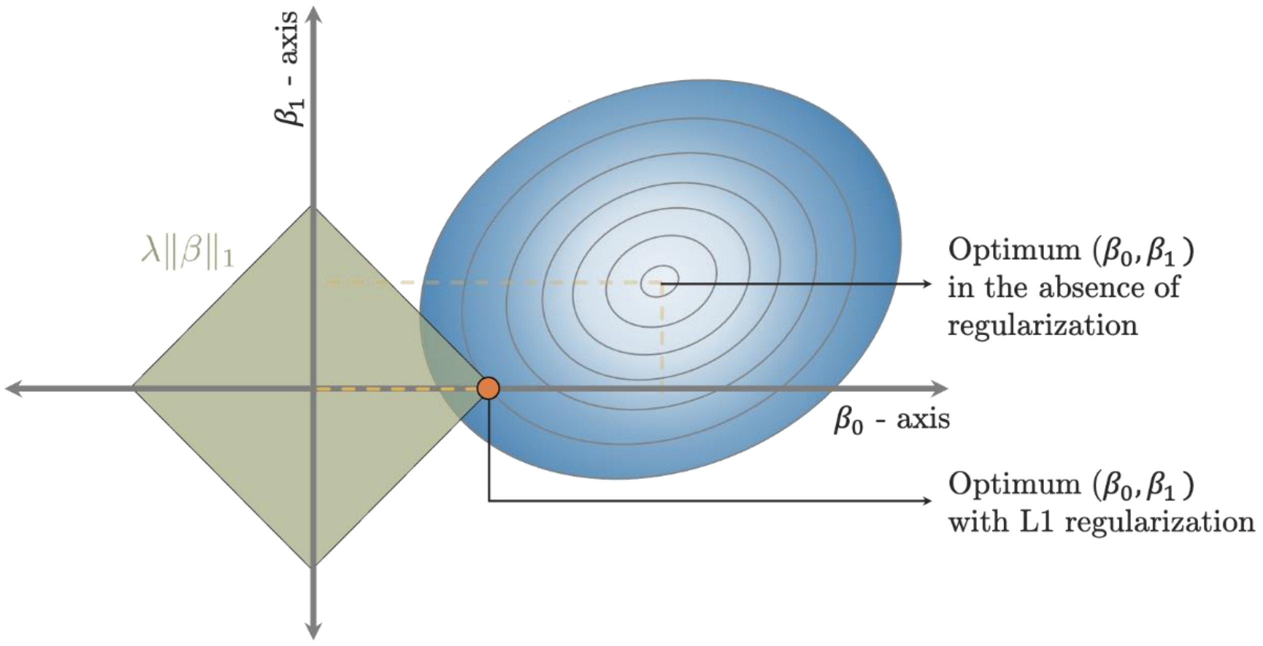


Fig.2: L1 regularization example.

NTNU

Source: Types of Regularization in Machine Learning | by Aqeel Anwar | Towards Data Science

L2 Regularization

L2 Regularization

- Ridge Regression
- L2-norm squared
- Parameters of bigger value are more effected

Ridge Regression Cost Function = Loss Function + $\lambda \sum_{\{j=1\}}^{m} \beta_j^2$

Controls the strength of regularization

Model parameters

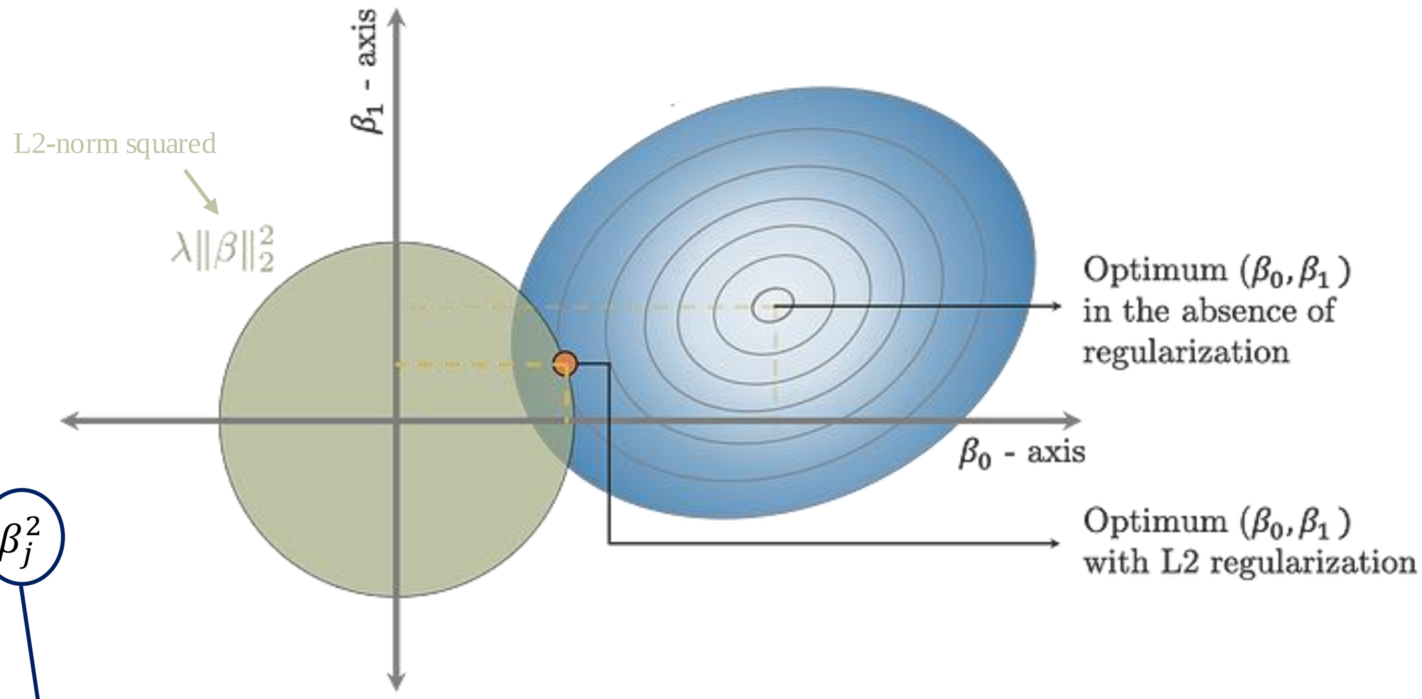


Fig.3: L2 regularization example.

Source: [Types of Regularization in Machine Learning | by Aqeel Anwar | Towards Data Science]

L2 Illustration

Let's make a unit "ball":

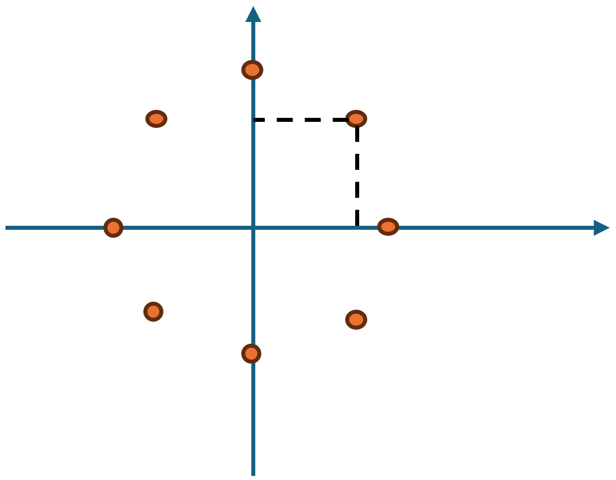
$$\sum_{\{j=1\}}^{m} \beta_j^2 = 1$$

$$\beta_1^2 + \beta_2^2 = 1$$

$$1^2 + 0^2 = 1$$

$$\left(\frac{1}{\sqrt{2}}\right)^2 + \left(\frac{1}{\sqrt{2}}\right)^2 = 1$$

…



L2 Regularization

Theoretical part

- Parameters of bigger value are more effected
- Not sparsifying
- Emphasize model's essential features, by making some coefficients (parameters) close to zero.

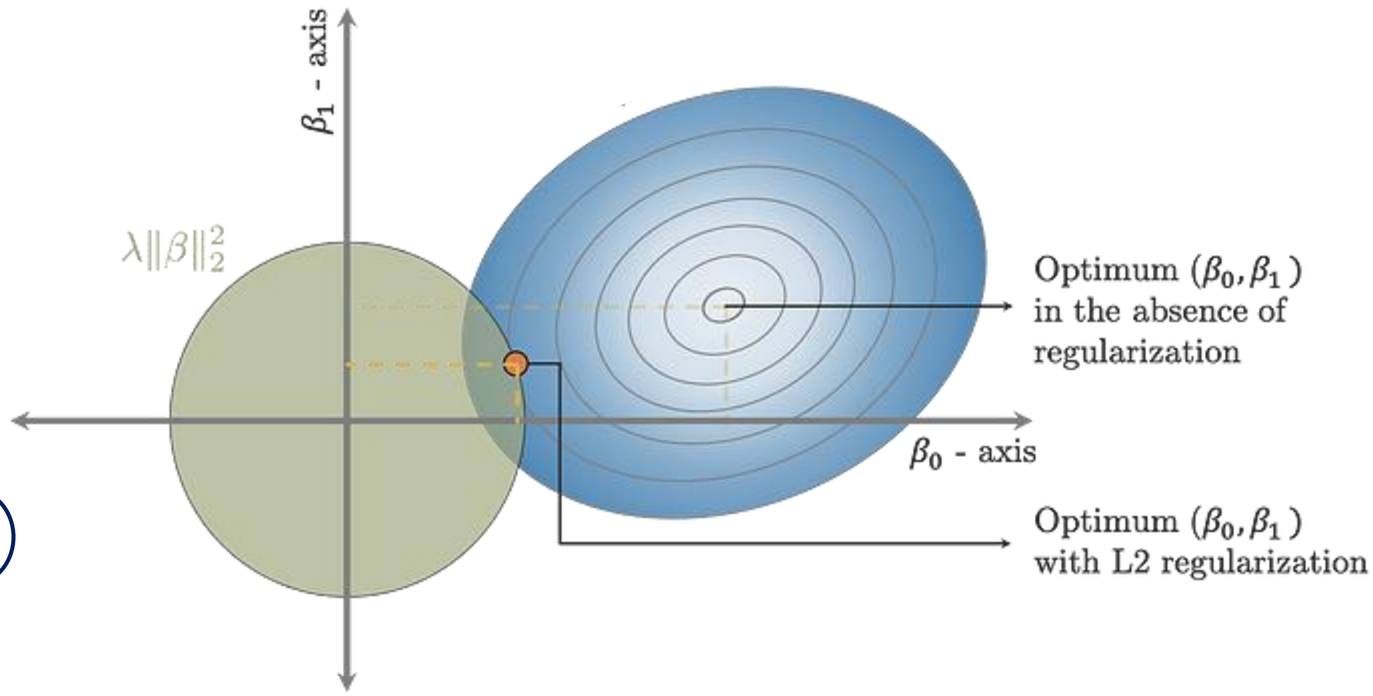


Fig.3: L2 regularization example.

Ridge Regression Cost Function = Loss Function + $\lambda \sum_{\{j=1\}}^{m} \beta_j^2$

Controls the strength of regularization

Model parameters

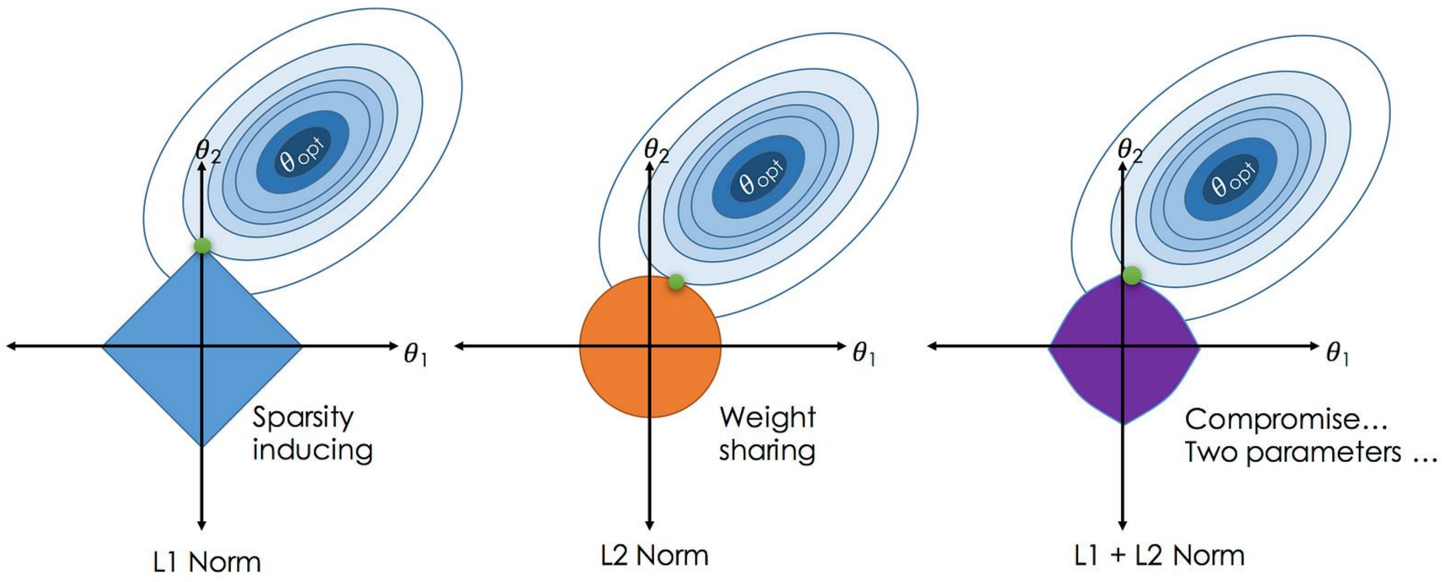
NTNU

Source: [Types of Regularization in Machine Learning | by Aqeel Anwar | Towards Data Science]()

Elastic net regression

- Combine l1 and l2 regularization
- Model with many useless variables -> use l-1
- Many useful variables –> use l-2

- Good when there are correlations between parameters

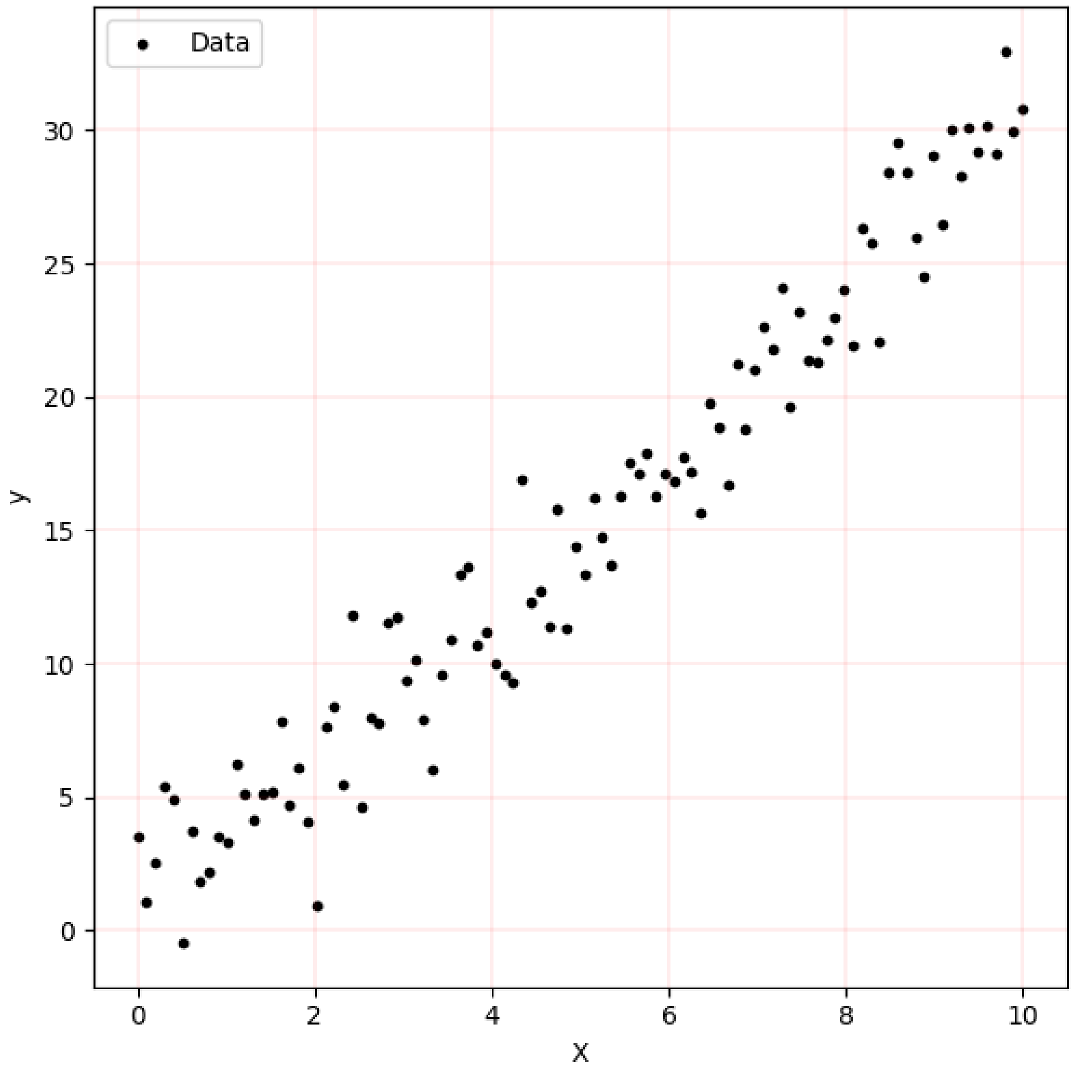


Elastic net regression Cost Function = Loss Function + $\lambda_1 \sum_{\{j=1\}}^{m} |\beta_j| + \lambda_2 \sum_{\{j=1\}}^{m} \beta_j^2$

Regression example

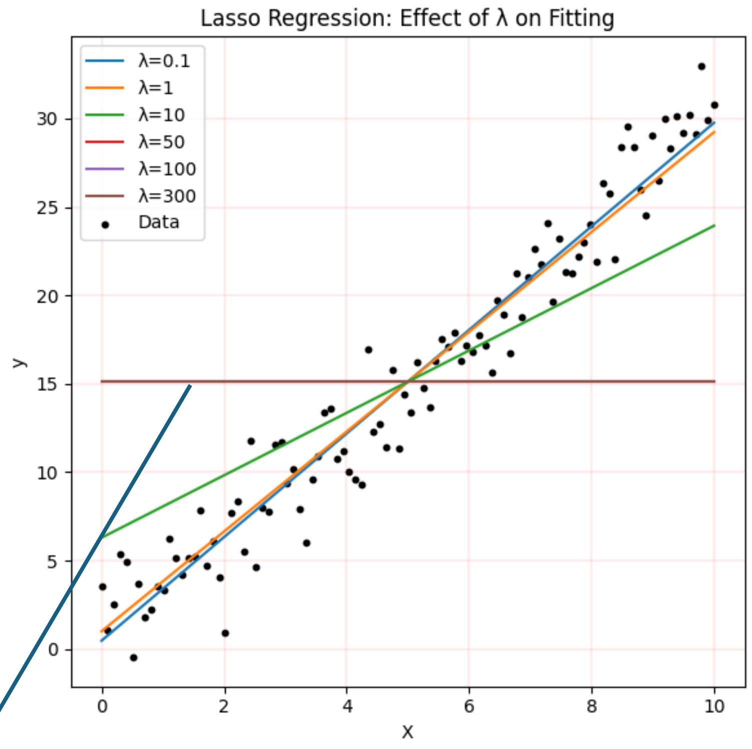
- Want to find line that fits the data
- Optimize for β
- See how the λ for L1 and L2 regularization affects the slope of the fitted line

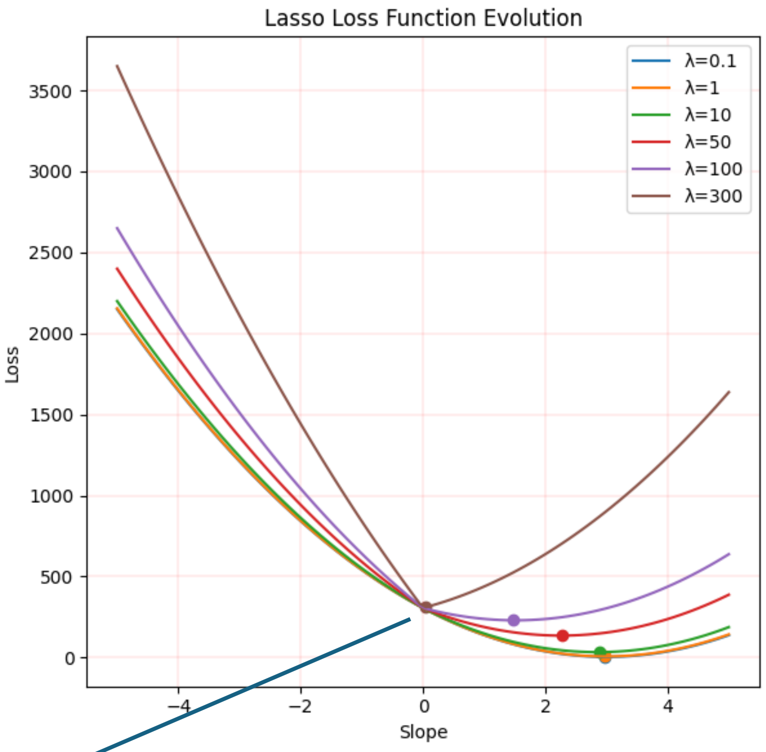
$$y = \beta x + c$$



L1

Results

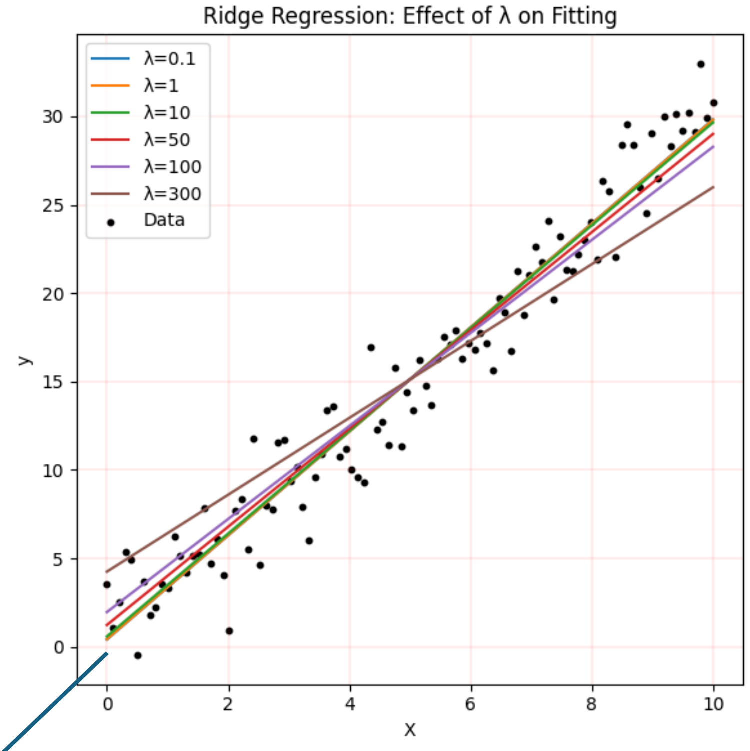


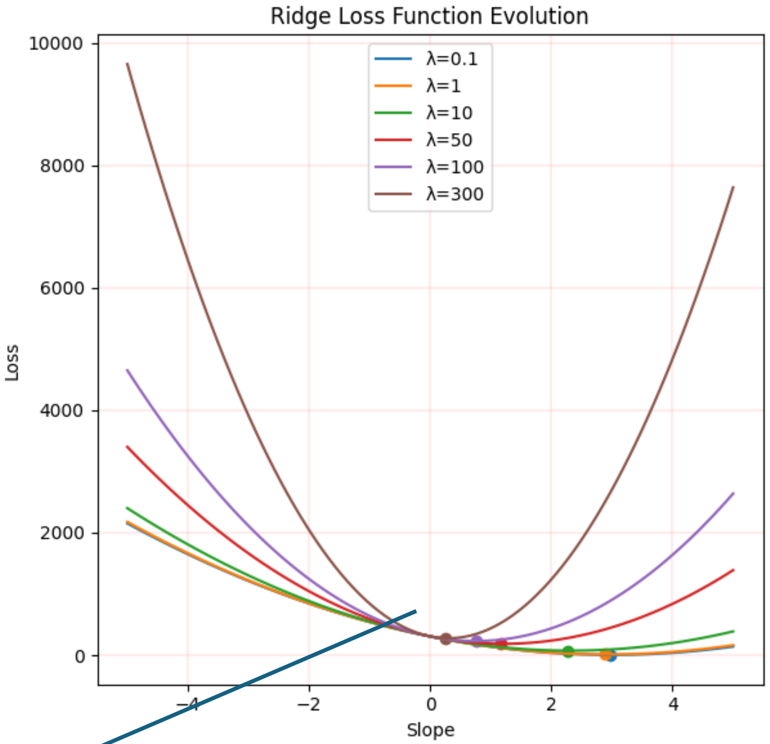


Observe as we increase the lambda, the minima of the loss function move towards slope = 0, and eventually become zero, the regression line is flat

L2

Results





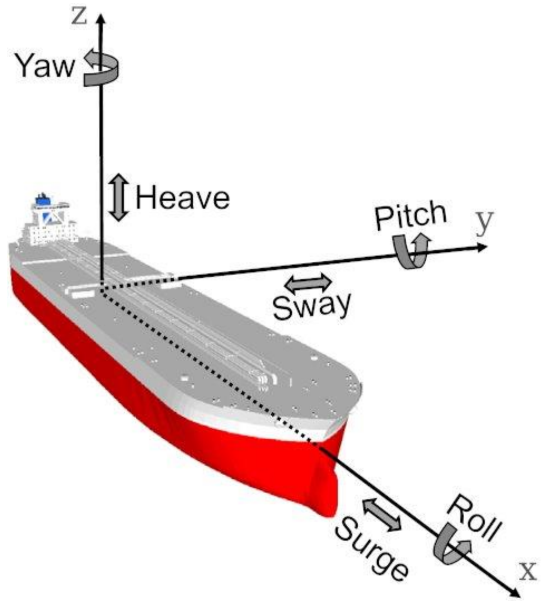
As lambda increases, minima is drawn asymptotically to slope = 0 → slope ≠ 0 → we preserve the variable x

NTNU

Another Example: L1 feature selection on real data sets

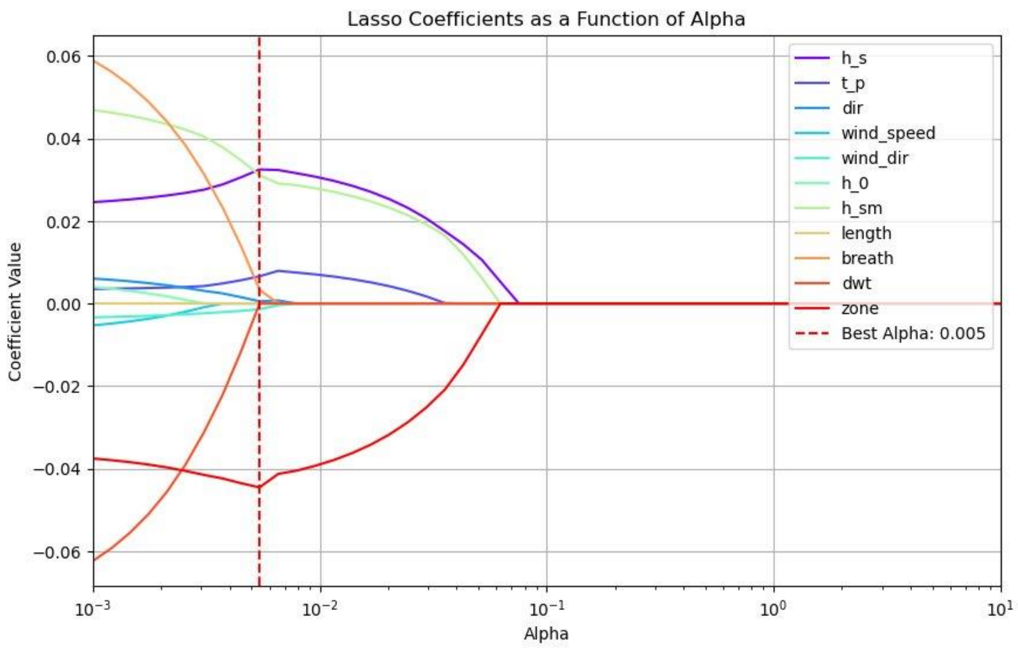
Results

- Ships movement dataset recorded in the Outer Port of Puntea Langosteira (A Coruña, Spain) from 2015 until 2020.
- Source: https://github.com/aalvarell/ship-movement-dataset
- Variables:
  - h_s (m): significant wave height.
  - t_p (s): peak wave period.
  - dir (deg): mean wave direction.
  - wind_speed (km/h): mean wind speed.
  - wind_dir (deg): mean wind direction.
  - h_0 (m): sea level with respect to the zero of the port.
  - h_sm (m): significant wave height measured by a tide gauge.
  - length (m): ship length.
  - breadth (m): ship breadth.
  - dwt (tonnes): deadweight tonnage
  - zone: . The port is divided into 12 berthing zones (the port operator provides it).
- Purposes:
  To study the influence of different predictors on the heave motion model for ship



Another Example: L1 feature selection on real data sets

Results



Notable results:

- significant wave, height h_s, has the highest correlation with heave movement y, which is to be expected.

- Some features that can be stripped (low correlation):
  - Mean wave direction, dir,
  - Sea level with respect to the port, h_0,
  - Peak wave period, t_p,
  - Ship length, length,
  - Mean wind speed, wind_speed,
  - Mean wind direction, wind_dir.

Choosing Between the methods

Discussion

- **Use L1 Regularization (Lasso) when:**
  - You have many features but expect only a few to be important
  - Feature selection is important, as L1 can reduce irrelevant features to zero
- **Use L2 Regularization (Ridge) when:**
  - You have multicollinearity in your data
- **Use Elastic Net when**
  - When you want benefits of both
    - Feature selection (L1)
    - Multicollinearity Handling (L2)

https://medium.com/@fernando.dijkinga/explaining-l1-and-l2-regularization-in-machine-learning-2356ee91c8e3

Thank you!

NTNU