Machine Learning
Theory & Practice

Module 3: **Linear Regression**

Lecture 2: **Performance Tuning for Linear Regression**

Lecture Outline

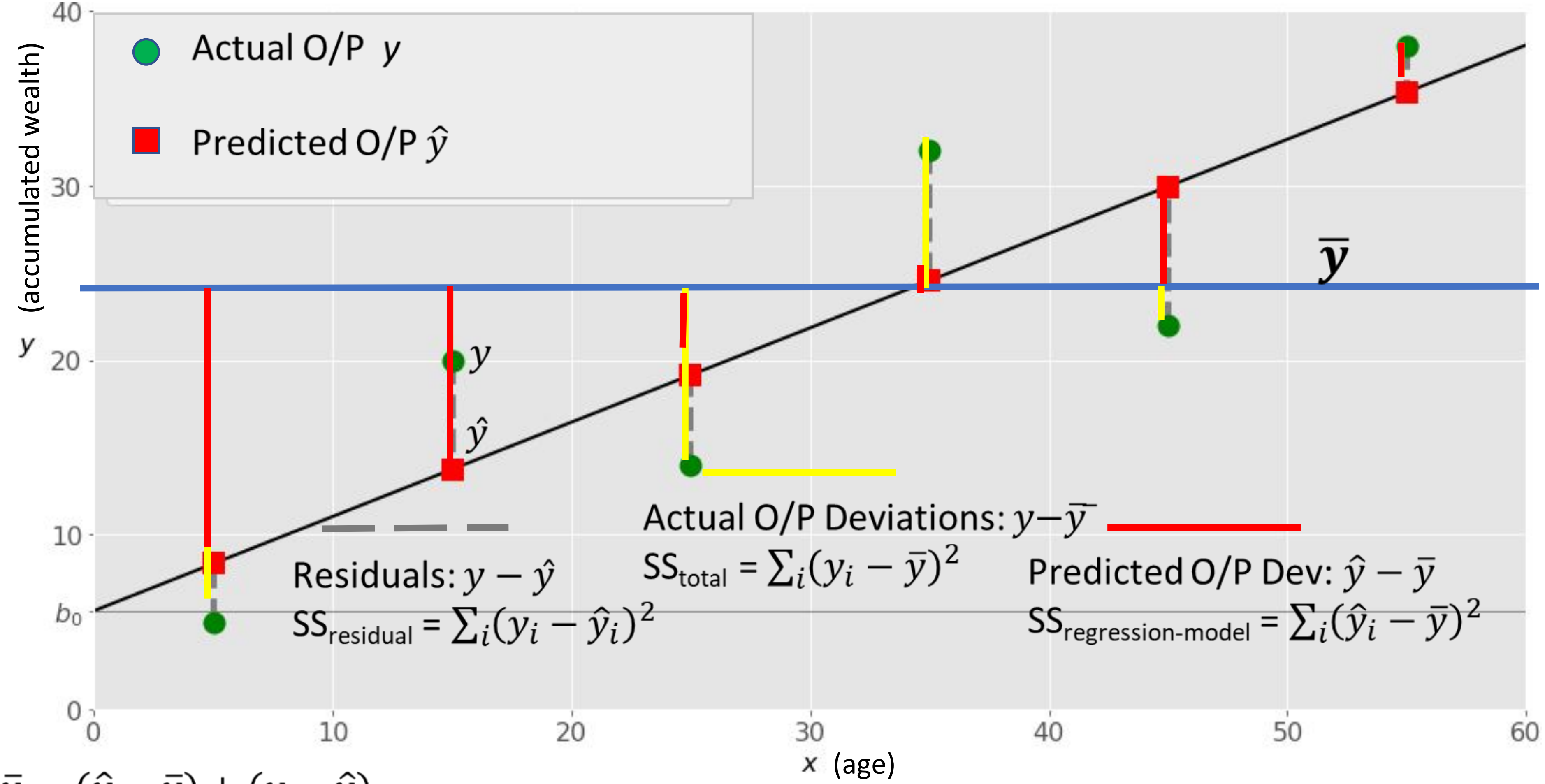
Topic 1: **R-squared and Adjusted R-squared**

Topic 2: **Regularization**

Topic 1: R-Squared and Adjusted R-Squared:

These are performance metrics that show how much of the output variation is explained by a LR model

Variation in Response



$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Coefficient of Determination: R-Squared - R^2

Variances:

- Total variance in actual response: $SS_{total} = \sum_{i=1...n}(y_i - \bar{y})^2$
- Error: $SS_{residual} = \sum_{i=1...n}(y_i - \hat{y}_i)^2$
- Variance in response given by Model: $SS_{reg\text{-}model} = \sum_{i=1...n}(\hat{y}_i - \bar{y})^2$
- $R^2 = \frac{SS_{reg\_model}}{SS_{total}}$: proportion of the variation in actual response that is explained by the LR model
- OR, $1 - \frac{SS_{residual}}{SS_{total}}$: 1- proportion of unexplained variation (error) in response

Issues with R-squared

- $$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_i (y_i - (w_0 + w_1 x_1 + \cdots))^2}{\sum_i (y_i - \bar{y})^2}$$

- It is biased: By adding many regressors, and by increasing training time, we get apparently better R^2
- May add insignificant predictors
- May start modelling noise
- Model overfits, and hence will be unable to generalize

Adjusted-R Squared $\bar{R}^2$

Mean Variances

- $\text{MS}_{\text{total}} = \frac{\sum_{i=1\ldots n}(y_i - \bar{y})^2}{n-1}$, n: No. of samples, n-1 : degree of freedom df

- $\text{MS}_{\text{regression-model}} = \frac{\sum_{i=1\ldots n}(\hat{y}_i - \bar{y})^2}{p}$, df=p, p= no of regressors (1 for SLR)
- $\text{MS}_{\text{residual}} = \frac{\sum_{i=1\ldots n}(y_i - \hat{y}_i)^2}{n-p-1}$, Note: $\text{df}_{\text{total}} = \text{df}_{\text{regression-model}} + \text{df}_{\text{residual}}$

- $\bar{R}^2 = 1 - \frac{\text{MS}_{\text{residual}}}{\text{MS}_{\text{total}}} = 1 - \frac{SS_{residual}}{SS_{total}} \frac{(n-1)}{(n-p-1)}$

Relationship between R^2 & $\bar{R}^2$

R^2 Versus $\bar{R}^2$

- LR: $R^2 = 1\text{-}SS_{res}/SS_{tot}$
- R^2 Always increases as more regressors are added.
- R^2 is biased estimate
- R^2 Does not penalize non-significant terms
- R^2 Is always positive
- R^2 is not suitable for statistical test of significance of weights

- LR: $\bar{R}^2 = 1\text{-}MS_{res}/MS_{tot}$
- $\bar{R}^2$ increases ONLY if an added regressor is significant.
- $\bar{R}^2$ is unbiased estimate
- $\bar{R}^2$ penalizes non-significant variables
- $\bar{R}^2$ can be negative, is always less than R^2
- $\bar{R}^2$ is suitable for statistical test of significance of weights

Topic 2: Regularization methods

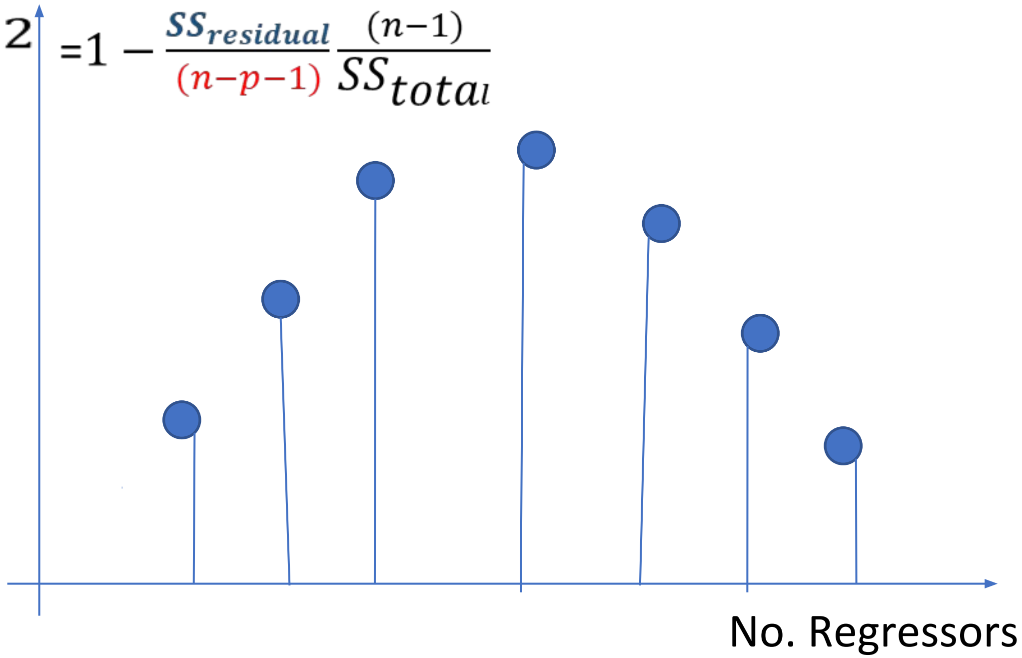
REGULARIZATION: PROCESS OF MAKING THE LEARNING MODEL SIMPLER

Regularization with Validation

- Make a hierarchy of regressors (X) in terms of their relative importance.
- Add Regressors, one/few at a time, in order of importance.
- Generate an optimized model for each set of regressors using cross validation, and monitor $\bar{\boldsymbol{R}}^{\boldsymbol{2}}$
- The point at which $\bar{\boldsymbol{R}}^{\boldsymbol{2}}$ reaches a maximum gives the ideal combination of regressors.

Example: Cost of House

$$\bar{R}^2 = 1 - \frac{SS_{residual}}{(n-p-1)} \frac{(n-1)}{SS_{total}}$$



Regressors:

1. **Floor Area**
2. **No of bedrooms**
3. **No of balconies**
4. **Type of locality**
5. **Green cover**
6. **Front door facing which direction**
7. **Educational background of neighbourhood**

Regularization with Modified Loss Functions

- Augment Ordinary Least Squares with regularization term:
    - LASSO Regression ⍰ L1 Regularization
    - Ridge Regression ⍰ L2 Regularization
    - Elastic Net Regularization

Least Absolute Shrinkage & Selection Operator(LASSO): L1 Regularization

Minimize cost function: 1) Ordinary Least Squares 2) Regularization Term

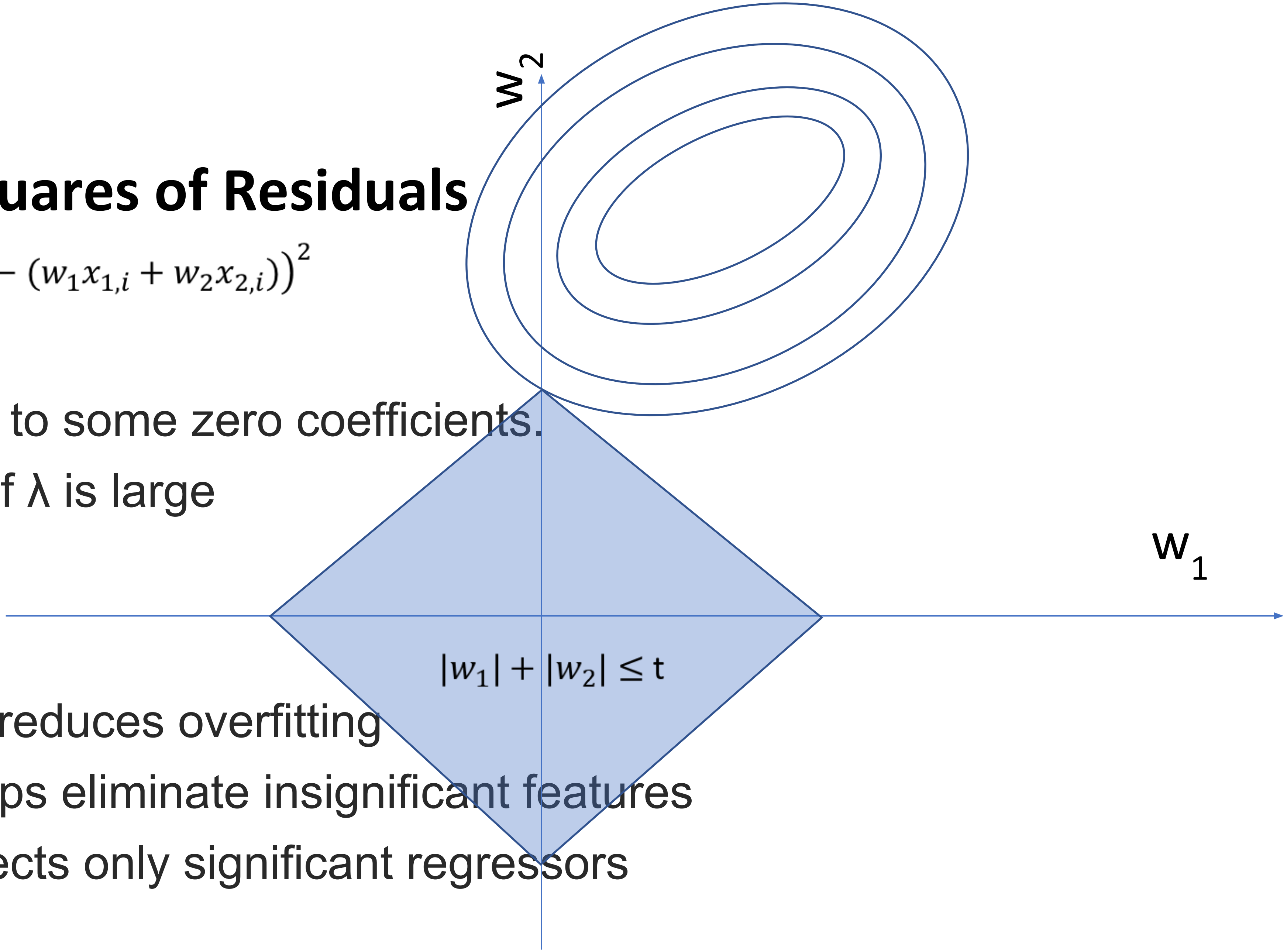
$$\textit{Minimize}\left\{ \boxed{\sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2} + \boxed{\lambda \sum_{j=0}^{p} |w_j|} \right\}$$

Forcing $\text{For some } t > 0,\ \sum_{j=0}^{p} |w_j| < t$

- L1 penalizes regressors by shrinking their weights
- Regressors that contribute little to error reduction are more penalized
- λ is the weighting factor for regularization to tune overfit ↔ underfit

Sum of squares of Residuals

$$\sum_i \left(y_i - (w_1 x_{1,i} + w_2 x_{2,i})\right)^2$$



- L1 can lead to some zero coefficients.
  especially if λ is large

- L1 not only reduces overfitting
  but also helps eliminate insignificant features
- LAS**S**O selects only significant regressors

Multicollinear Regressors

- The following regressors are highly positively correlated and each apparently impacts property cost.
  - Total Area
  - Number of rooms
  - Free areas: balconies and corridors
  - Size of rooms
- Without regularization, all coefficients would be inflated
- L1 retains only one of them and eliminates the rest

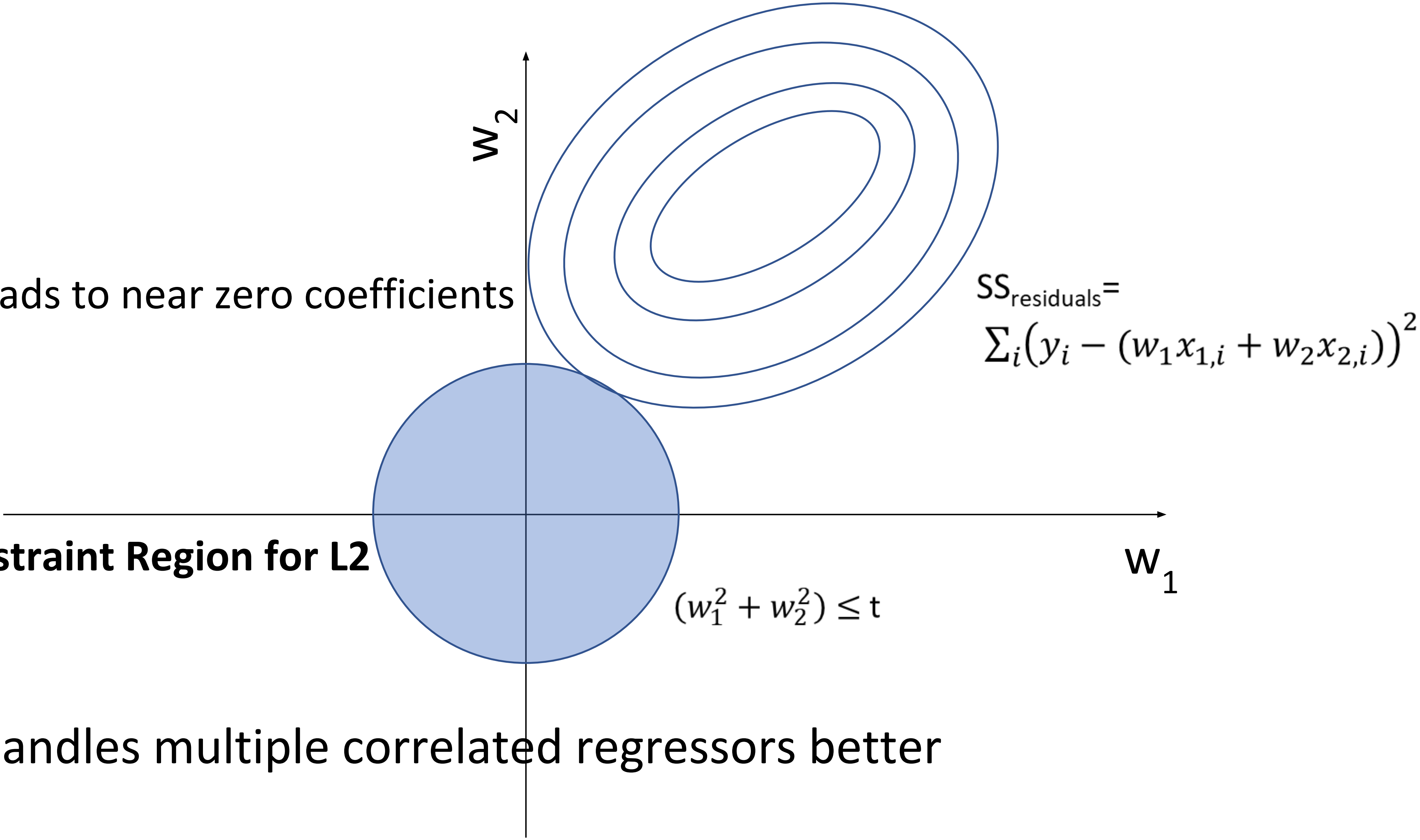
Ridge Regression : L2 Regularization

Minimize cost function: 1) Ordinary Least Squares 2) Regularization Term

$$\textit{Minimize}\left\{ \boxed{\sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2} + \lambda \boxed{\sum_{j=0}^{p} w_j^2} \right\}$$

Forcing, For some $c > 0$, $\sum_{j=0}^{p} w_j^2 < c$

- L2 leads to near zero coefficients

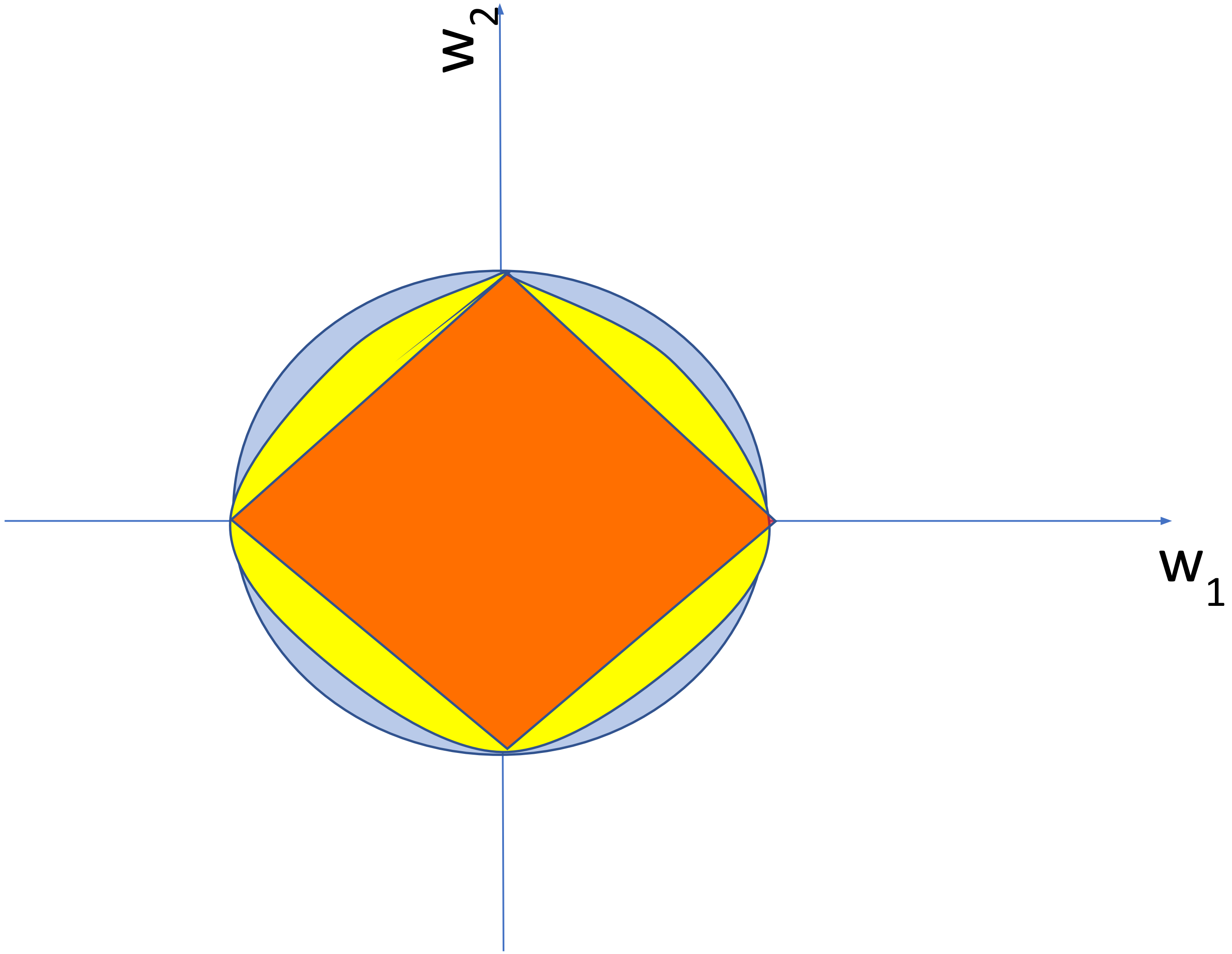


- L2 handles multiple correlated regressors better

Elastic Net Regularization

- Combines L1 and L2 Regularization
- Each has its own weighting factor
- $OF = Minimize\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m} w_j x_{j,i}\right)^2 + \lambda_1 \sum_{j=0}^{m}\left|w_j\right| + \lambda_2 \sum_{j=0}^{m} w_j^2\right\}$
- λ_1 and λ_2 allow:
  - A balance of attribute elimination ability and handling multiple correlated regressors
  - A proper tuning of overfitted model (both 0) to underfitted model (both large)

Constraint Regions in Ridge, LASSO & Elastic Net Regularization



Recap

- Adjusted Coefficient of Determination $\bar{R}^2$ indicates the proportion of normalized variation in response that is explained in an unbiased manner, by a LR model
- LASSO regression employs L1 regularization by adding "sum of absolute weights" term in Cost function, with weighting factor λ
- If λ is 0, there is no regularization and the model may overfit giving poor generalization
- If λ is large, most weights shrink and the model may underfit

Recap

- Ridge regression employ L2 regularization by adding “sum of squares of weights” to OLS, with weight factor λ
- L2 regularization cannot eliminate any regressor but can appropriately shrink insignificant ones
- Ridge regression handles multiple correlated features by diminishing or enhancing them simultaneously, instead eliminating all but one (as in L1)
- Elastic Net regularization employs a mix of L1 and L2 with their own weighting factors to create a balanced model

Balance is a sense of harmony….