機械学習特論

~ 理論とアルゴリズム ~

(Regularized regression: ridge and lasso)

講師：西郷浩人

Review: Ordinary Least Squares Regression (OLS)

- Least squares regression is often denoted exchangeably as OLS, Linear model or linear regression.
- The objective function to minimize is

$$min_{\boldsymbol{\beta}} ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}||^2$$

- Analytic solution is available as

$$\boldsymbol{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

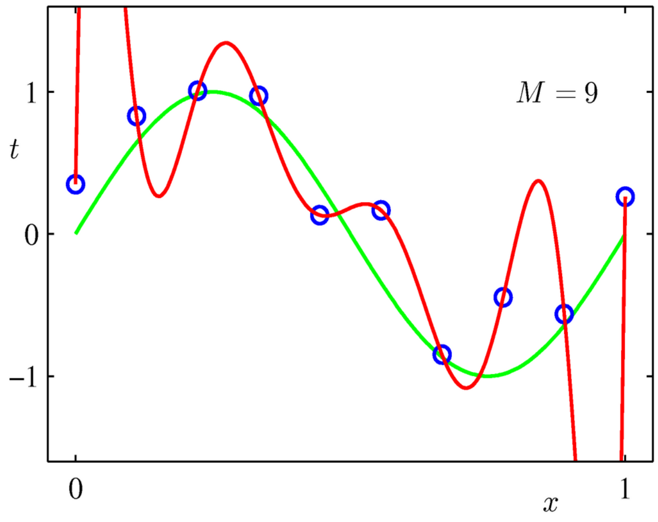
Prev. ex.2 Non-linear model

- Which of the following model fits best to the data in the table ?
  - Linear model $y=\beta_0+\beta_1 x$
  - Quadratic model $y=\beta_0+\beta_1 x+\beta_2 x^2$
  - Cubic model $y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$
- Compare the three models in terms of R^2, and answer the question. Next slide provides hints.

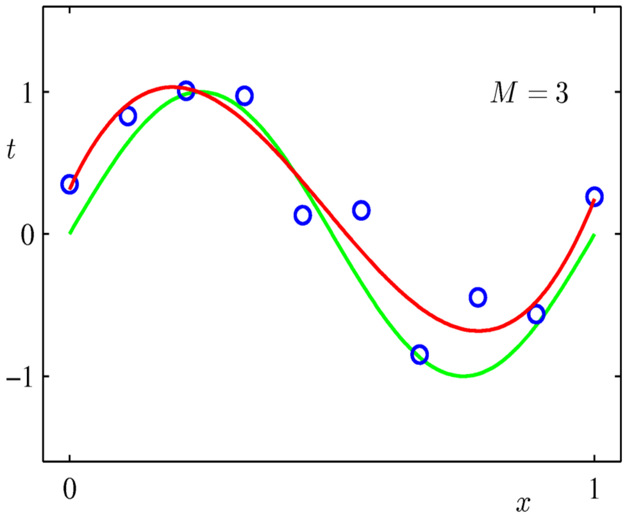
x	600	700	800	950	1100	1300	1500
y	253	337	395	451	495	534	573

Regularization in regression

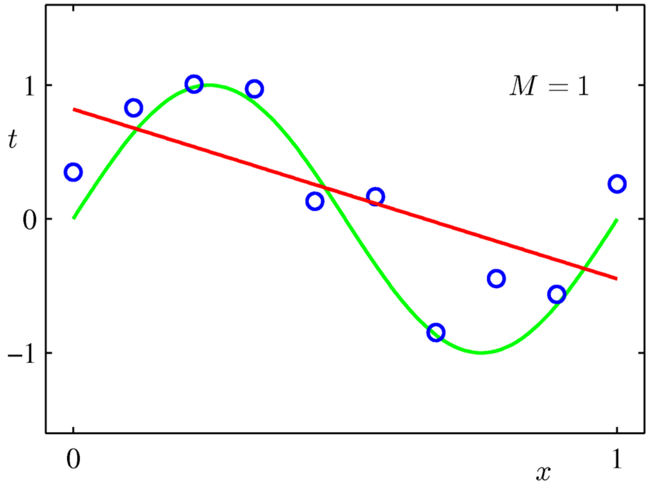
- In general, one can fit regression model to data as much as possible by using more variables or higher order polynomial models. But it often leads to degradation in fitting performance to future unknown data.
- The ideal rule is considered to lie in between overfitting(過適合) and underfitting.
- How can we tune between the two extremes ?



Overfitting



Optimal



Underfitting

© Bishop, Pattern Recognition and Machine Learning, 2007, Springer

Introducing Regularizers（正規化項） to Ordinary Least Squares

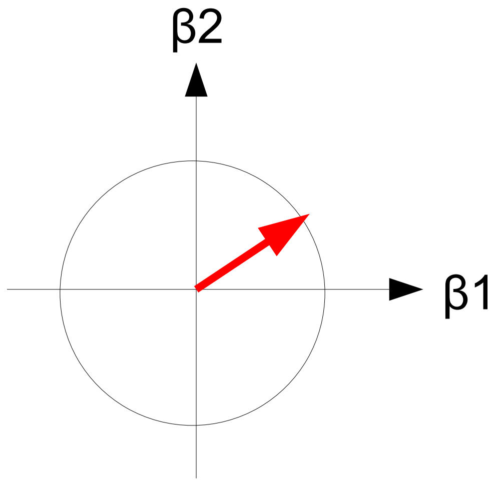
Introducing Regularization to OLS

- The objective function to minimize is

$$min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - y\|^2$$

- There is no constraint on **β**, so it can take infinitely large values.
- Constraints on **β** required ?
  - E.g., place it on an unit radius circle.

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



- New objective

$$min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - y\|^2$$
$$s.t. \|\boldsymbol{\beta}\| = 1$$

Vector Norms

- definition $$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

  - p =1: 1-norm (manhattan distance)

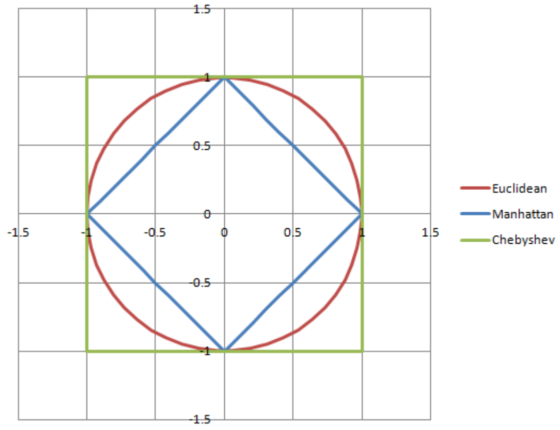
    $$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

  - p = 2: 2-norm (euclidian distance)

    $$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

  - p = ∞ : infinity-norm (chebyshev distance)

    $$\|\boldsymbol{x}\|_\infty = max_i |x_i|$$



Ridge regression

$$min_{\beta}\,\lambda\,\|\boldsymbol{\beta}\| + \|\boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{y}\|^{2}$$

- Minimizes *RSS* while placing **β** on a hyper sphere.
- 2-norm(L2) regularization on the coefficient vector **β**.
- Analytic solution available as

$$\boldsymbol{\beta}^{\boldsymbol{ridge}} = \left(\boldsymbol{X}\,'\,\boldsymbol{X} + \lambda\,\boldsymbol{I}\right)^{-1}\boldsymbol{X}\,'\,\boldsymbol{y}$$

OLS

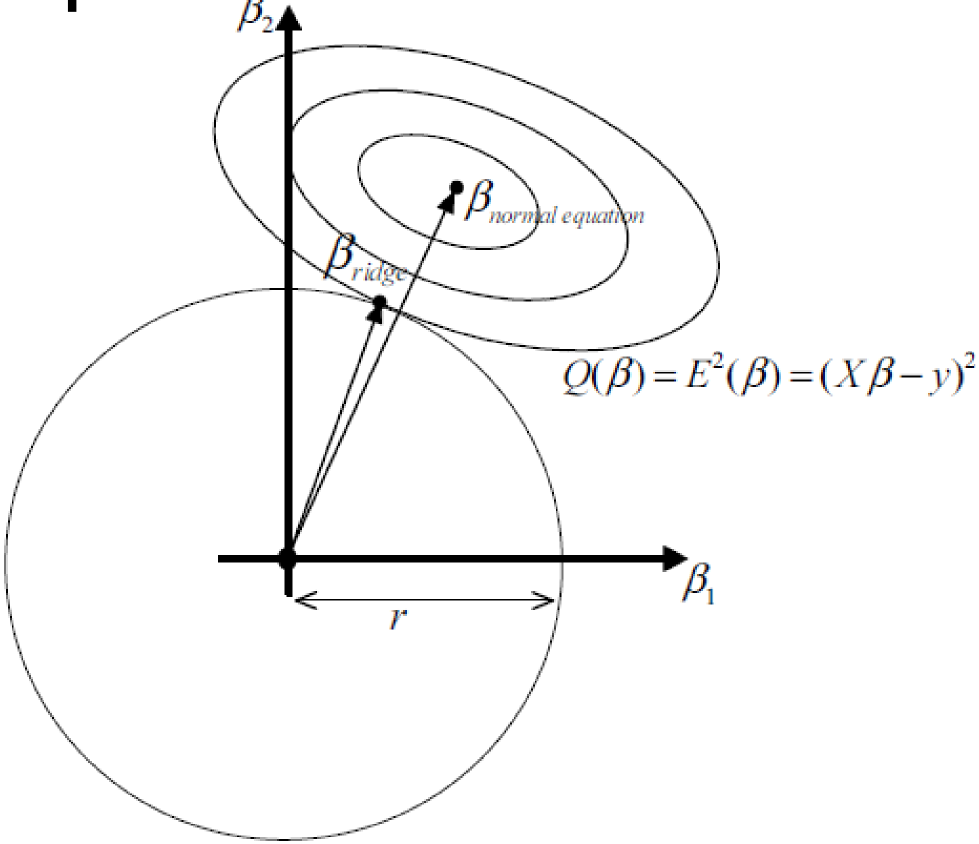
$$\boldsymbol{\beta}^{\boldsymbol{OLS}} = \left(\boldsymbol{X}\,'\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}\,'\,\boldsymbol{y}$$

  - λ is a regularization parameter.

- Bigger λ makes **β** closer to each other.
- Bigger λ can make the matrix $X^{T}X+ \lambda I$ apart from rank defficient, i.e., more numerically stable.

Solution of Ridge regression

$$min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\| + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$



β_2
$\beta_{normal\ equation}$
β_{ridge}
$Q(\beta) = E^2(\beta) = (X\beta - y)^2$
β_1
r

$$\boldsymbol{\beta}^{ridge} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

Role of regularization

- In ridge regression, changing λ corresponds to changing the radius of a circle on which β lies.
- When λ=0, solution of ridge regression (RR) coincides with that of ordinary least squares. As λ increases, the solution of RR gets far from the OLS solution. When λ is large, the objective function concentrates on laying β on a circle rather than minimizing the sqaured error (RSS).

(Ex.1)

$$min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\| + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$

- Show that the solution of the equation above is obtained as

$$\boldsymbol{\beta}^{ridge} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

Hint: Let the objective function be L, then take a derivative w.r.t. β and set it to zero.

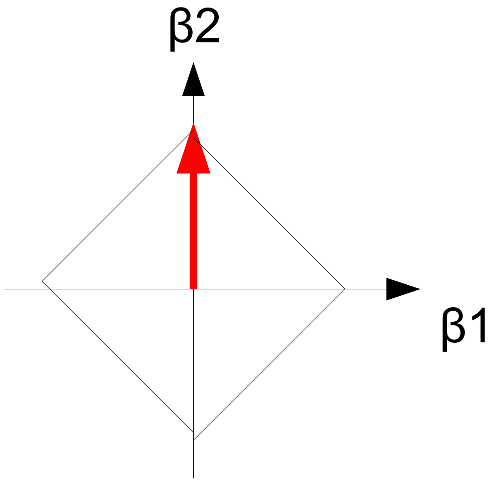
LASSO regression

$$min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$

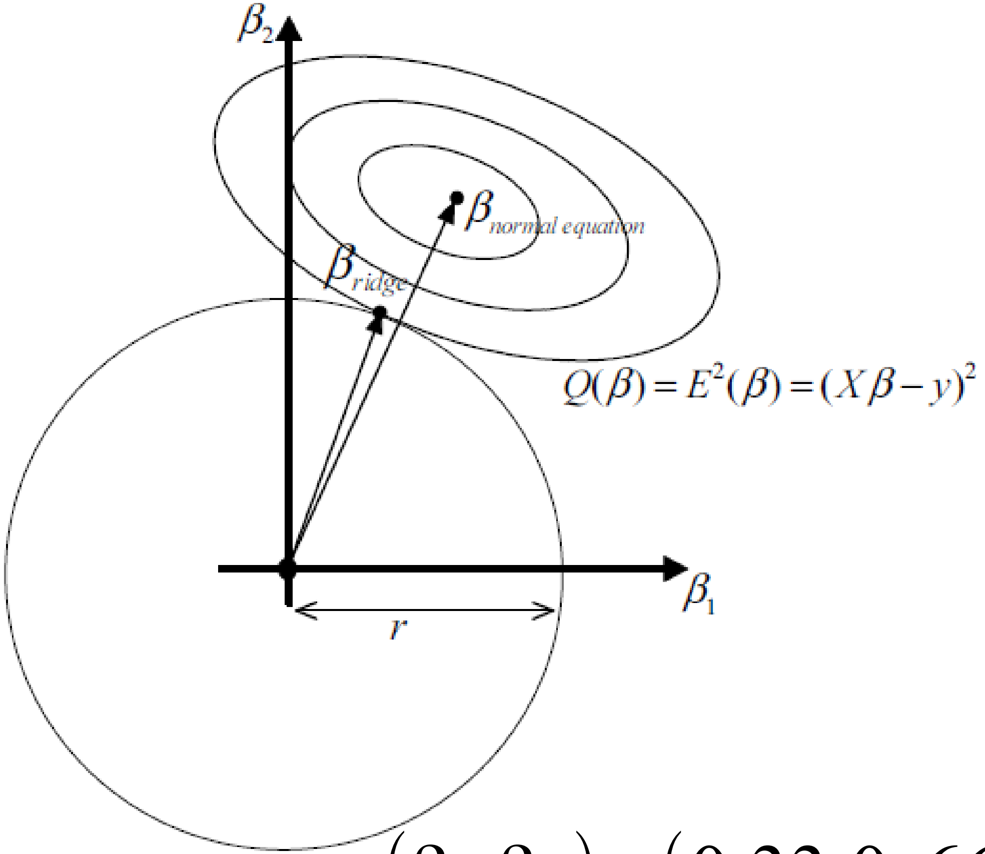
- Minimizes RSS while placing **β** on a hyper polygon.
- 1-norm(L1) regularization on a coefficient vector.
- Solving this problem had been nontrivial. In the original LASSO literature, the solution was found by solving quadratic programming (Tibshirani 1996).

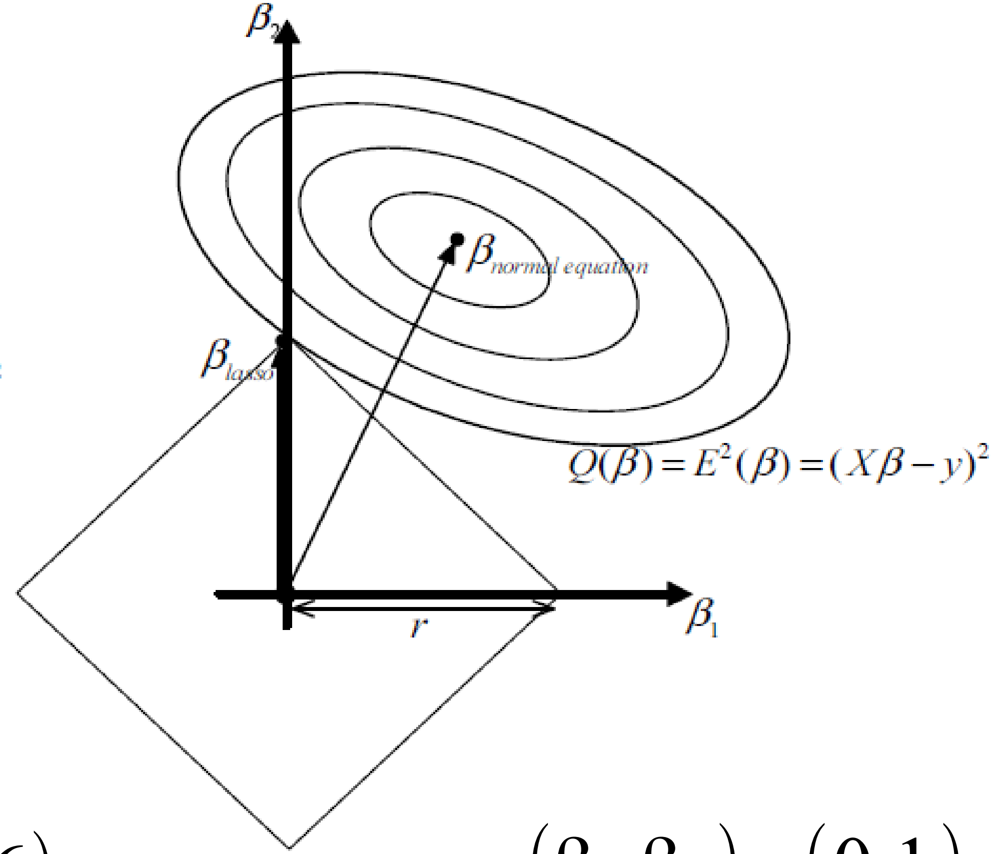
Ridge

$$min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$



L2(Ridge) vs L1(LASSO)


β_2
$\beta_{normal\ equation}$
β_{ridge}
$Q(\beta) = E^2(\beta) = (X\beta - y)^2$
β_1
r


β_2
$\beta_{normal\ equation}$
β_{lasso}
$Q(\beta) = E^2(\beta) = (X\beta - y)^2$
β_1
r

$(\beta_{1,}\beta_2)=(0.33,0\,.66)$

$(\beta_{1,}\beta_2)=(0,1)$

The contour shows the region with the same training error. We can easily obtain OLS solution on training data, but that might overfit to the training data, so we move it away (regularization).

Ex.2: Ridge vs LASSO

Load crime.txt, and compare ridge regression with lasso regression.

Download today's data, and try the following.

```
init
lambda=0;
ridge_vs_lasso(X,y,lambda)
```

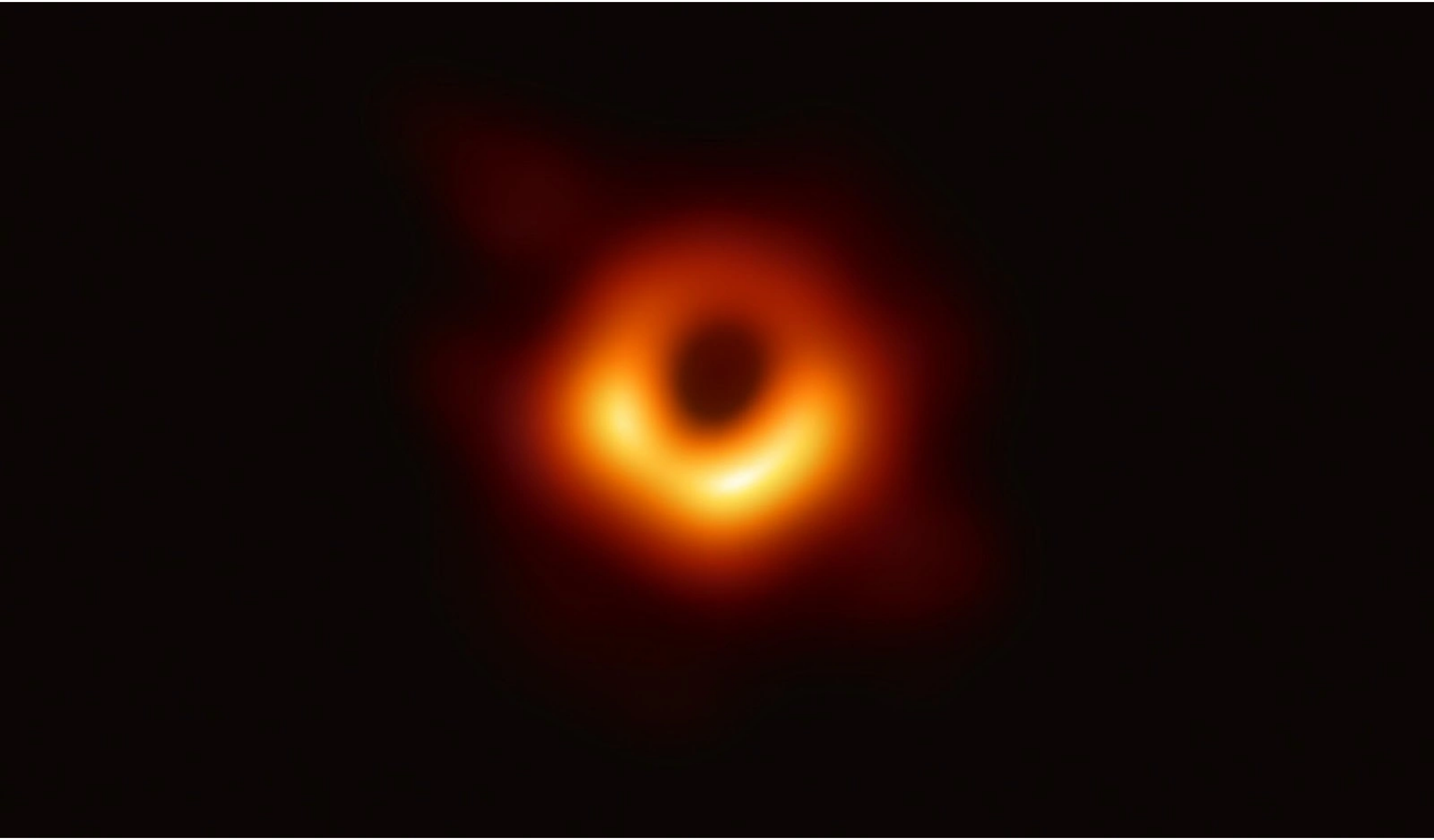
Change λ in the range {0, 0.1, 1, 10, 100}, and observe how RSS, sparsity and norm size changes.

- %Crime data for 50 U.S cities.
- %Y1 = total overall reported crime rate per 1 million residents
- %Y2 = reported violent crime rate per 100,000 residents
- %X3 = annual police funding in $/resident
- %X4 = % of people 25 years+ with 4 yrs. of high school
- %X5 = % of 16 to 19 year-olds not in highschool and not highschool graduates.
- %X6 = % of 18 to 24 year-olds in college
- %X7 = % of people 25 years+ with at least 4 years of college

What’s so cool about sparsity ?

- L1-norm is known to produce sparse coefficient weights. Which is..
  - Easily interpretable
    - Fewer variables are easier to interpret.
  - (Robust)
    - Fewer variables are robust against outliers.

Visualization of a black hole



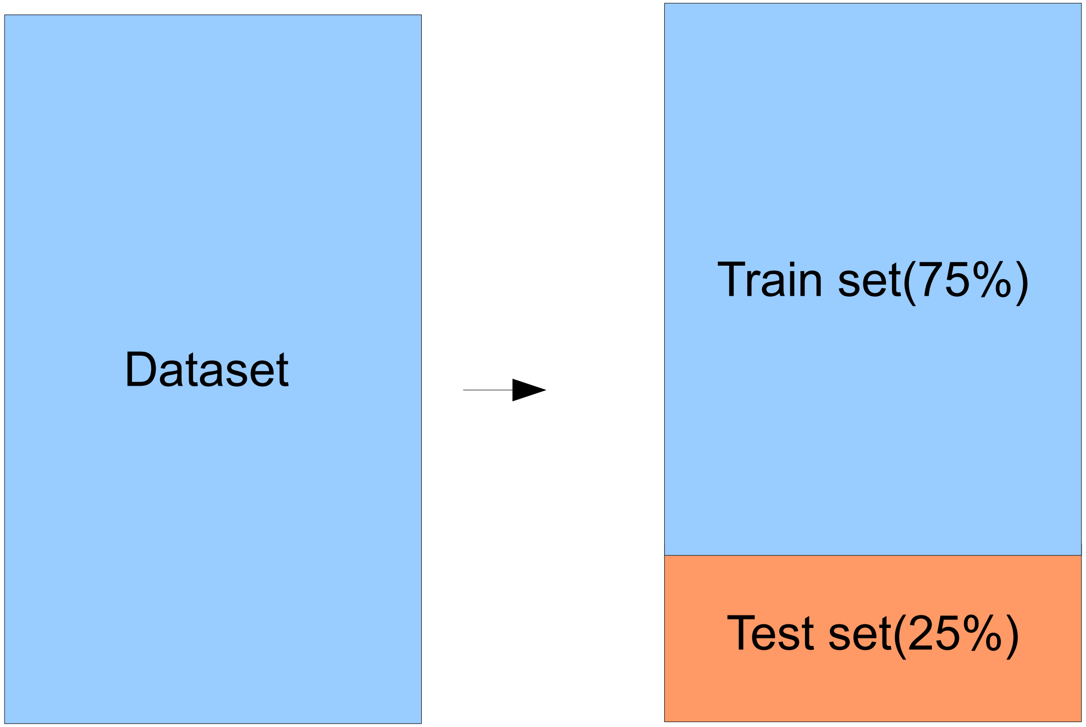
How to choose regularization parameter λ ?

- Formally, we **should not use training data** for choosing the regularization parameter, since it leads to overfitting.

- Following two methods are often employed
  - Use Validation Set
    - not good enough performance
  - Cross Validation
    - Often time consuming

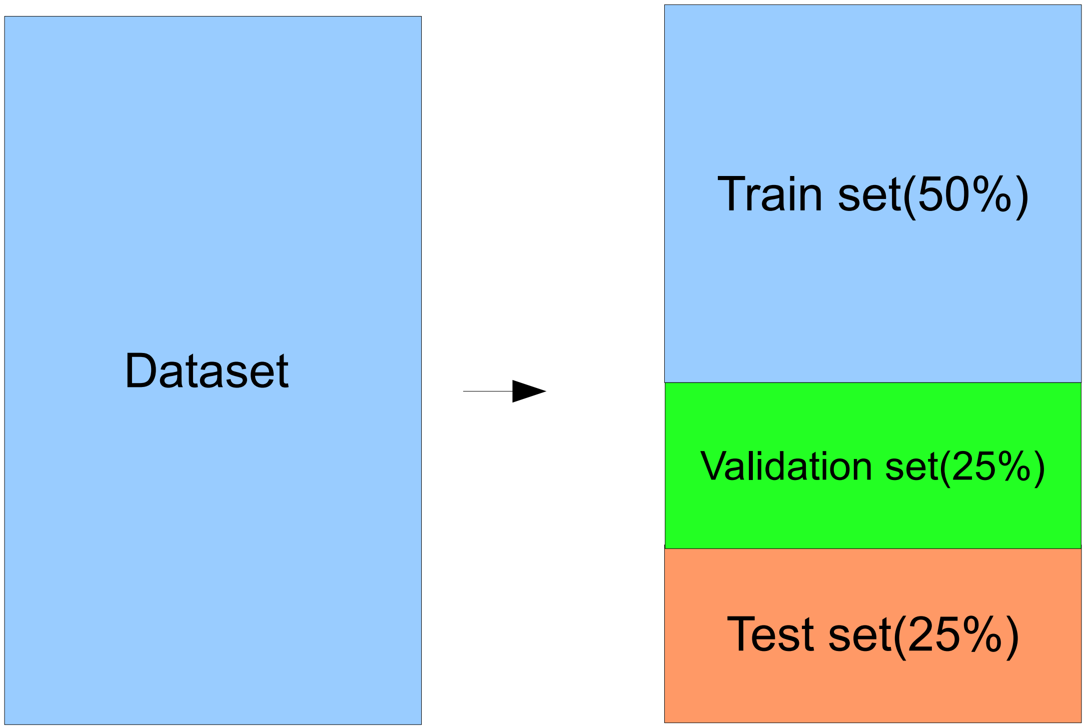
Standard situation

- Split data set into training and test.
- Learn coefficients from train set, then report performance in test set.



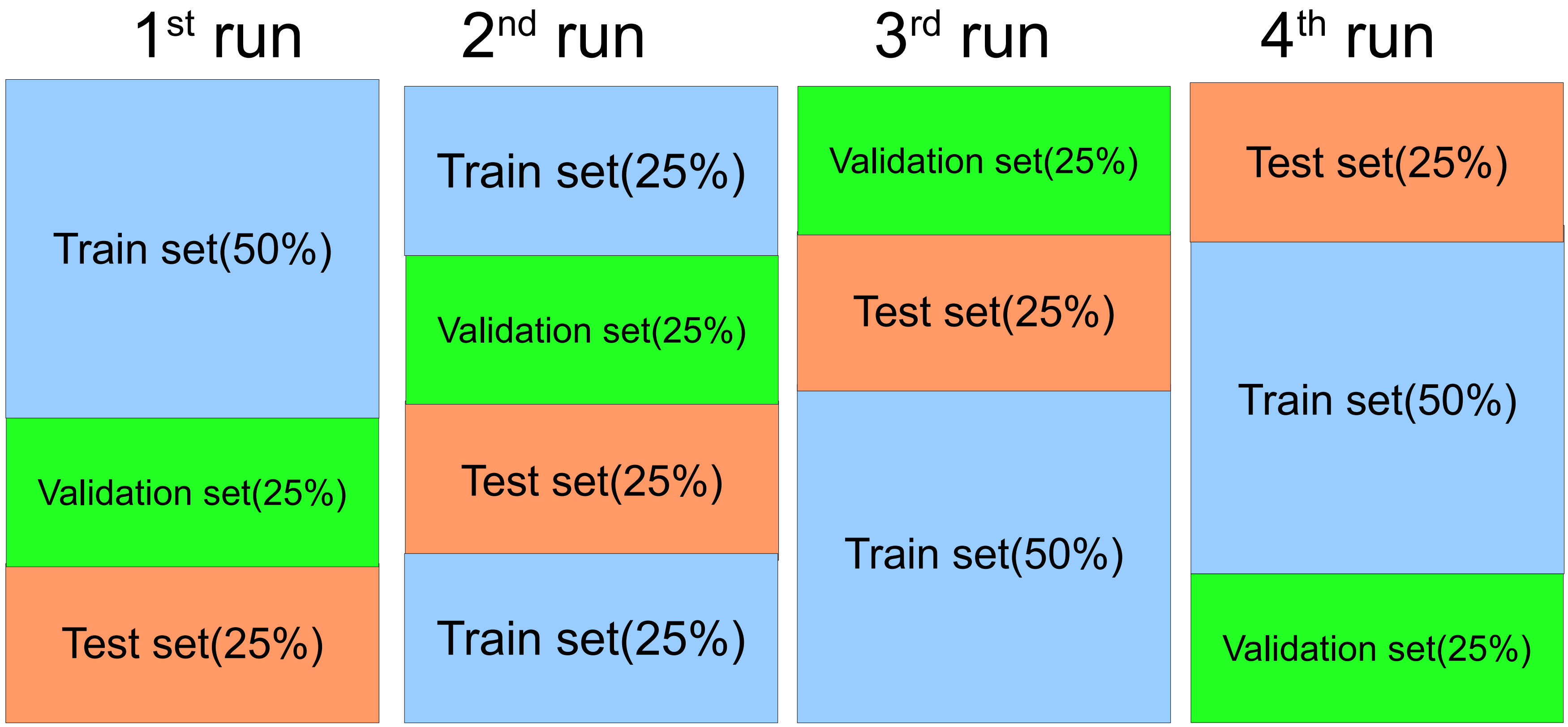
Validation set

- Split train set into “train” and “validation”.
- Choose regularization parameter that performs best in validation set, then report its performance in test set.



k-fold Cross Validation

- Split data into k-parts, leave one set out and train using remaining (k-1) sets, evaluate on the left-out test set.
- The following is an example of 4-fold cross validation using validation set.
- If we want to try different set of regularization parameters, this method could be extremely slow.



1st run

Train set(50%)

Validation set(25%)

Test set(25%)

2nd run

Train set(25%)

Validation set(25%)

Test set(25%)

Train set(25%)

3rd run

Validation set(25%)

Test set(25%)

Train set(50%)

4th run

Test set(25%)

Train set(50%)

Validation set(25%)

Leave-one-out cross validation

- Same as k-fold cross validation, but here k is the number of al the data.
- Often abbreviated as LOOCV.

(Ex. 3): model selection by cross-validation

- Split the crime data for 2-fold cross validation, and Report the mean R-squared from test datasets.
  - In each fold, choose a regularization parameter from {1,100,10000}.

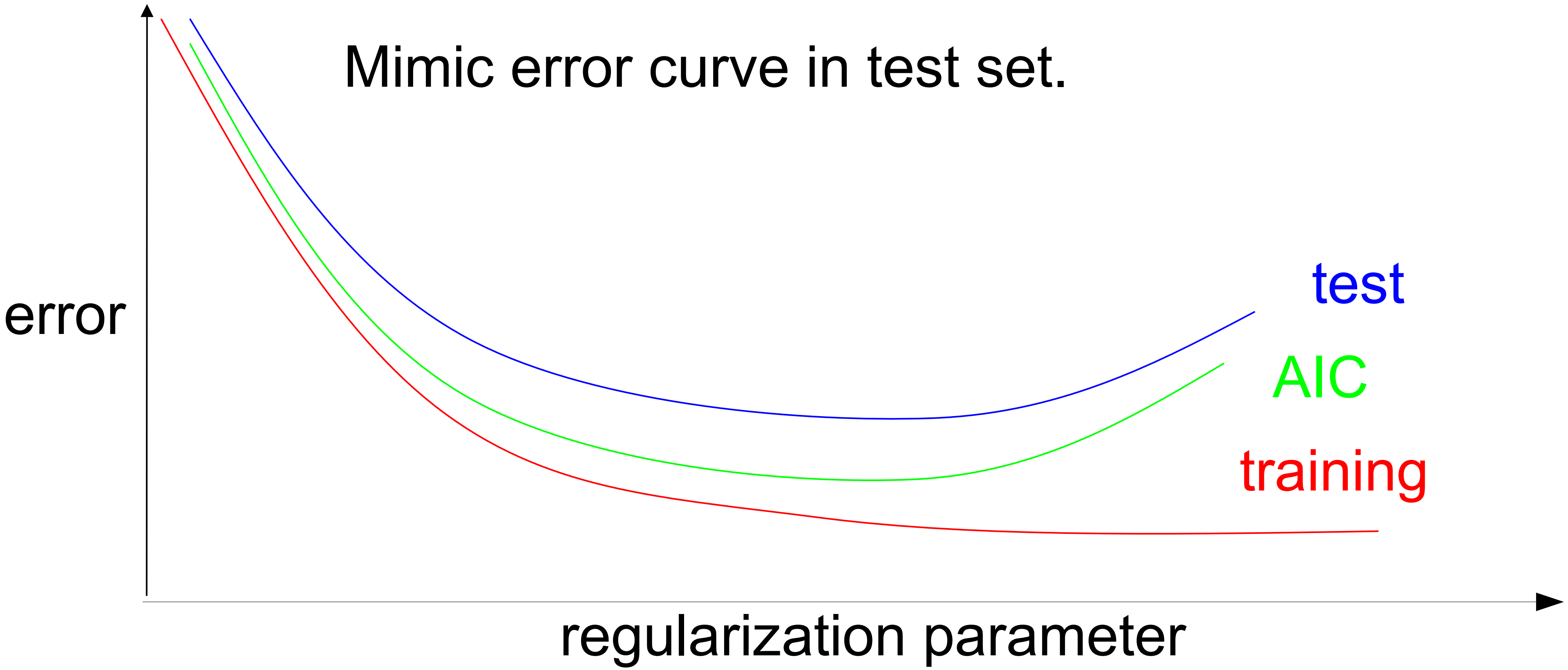
Information criterion （情報量基準）

~model selection based on training data~

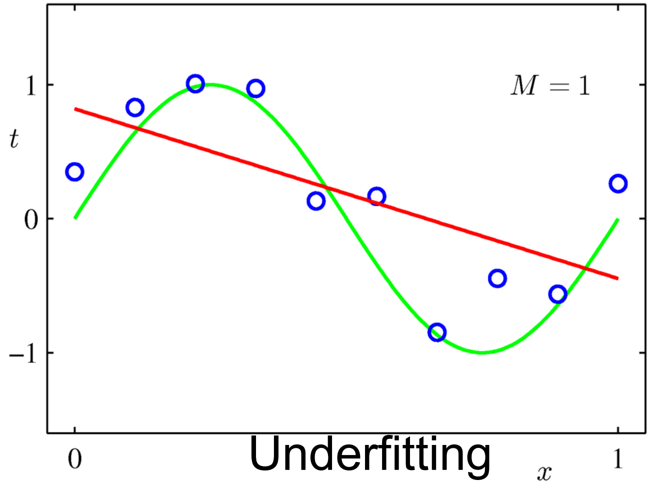
- Akaike Information Criterion (AIC)
  - = -2 loglikelihood + 2 p
    - where p is the number of coefficients
  - penalize the model with more variables (p).
- Bayes Information Criterion (BIC)
  - = -2 loglikelihood + p log(n)
    - where n is the number of data points.
  - Also known as minimum description length (MDL)

BIC assigns more weights on the number of parameters than AIC does. AIC is preferred for rather small data, while BIC is preferred for rather big data.

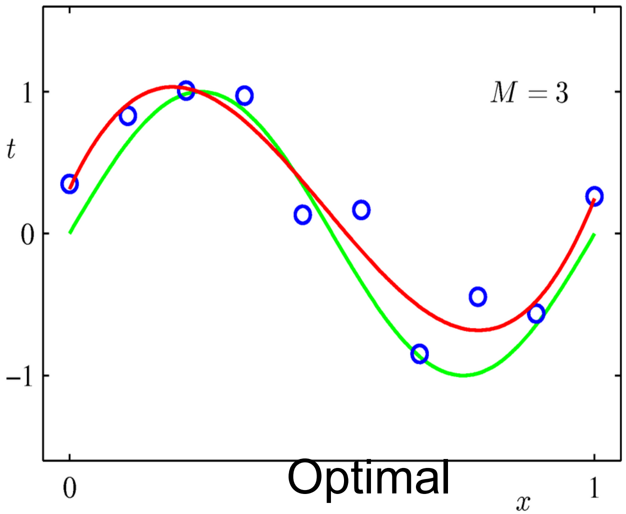
Model selection with AIC



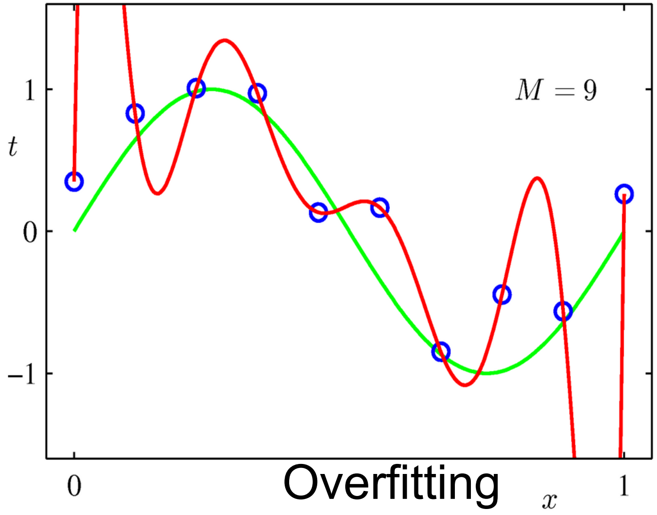
Mimic error curve in test set.



Underfitting



Optimal



Overfitting

AIC in least squares regression

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- The least squares solution above maximizes likelihood function below.

$$\log L(\boldsymbol{\beta}, \sigma^2) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{x_i}'\boldsymbol{\beta} - y_i)^2$$

  - The first term on the right hand side of the equation is a constant, and the 2nd and the 3rd term can be obtained from RSS, since

$$\sigma^2 = \frac{1}{n}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}\beta_j - y_i\right)^2 = \frac{RSS}{n}$$

$$AIC = -2\,logL + 2\,p = n\log(2\pi) + n\log\left(\frac{RSS}{n}\right) + n + 2\,p$$

$$\approx n\log\left(\frac{RSS}{n}\right) + 2\,p$$

  - where p is the number of parameters in the model

Prev. ex.2 Non-linear model

- Which of the following model fits best to the data in the table ?
  - Linear model $y=\beta_0+\beta_1 x$
  - Quadratic model $y=\beta_0+\beta_1 x+\beta_2 x^2$
  - Cubic model $y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$
- Compare the three models in terms of R^2, and answer the question. Next slide provides hints.

x	600	700	800	950	1100	1300	1500
y	253	337	395	451	495	534	573

```
>X=[600   700   800   950  1100  1300  1500]';
>y=[253  337  395  451  495  534  573]';

% linear model
>X=[ones(7,1) x];
>beta=X¥y; r2=1-sum((X*beta-y).^2)/var(y)
0.55821

% quadratic model
>X=[ones(7,1) x x.^2];
>beta=X¥y; r2=1-sum((X*beta-y).^2)/var(y)
0.94204

% cubic model
>X=[ones(7,1) x x.^2 x.^3];
>beta=X¥y; r2=1-sum((X*beta-y).^2)/var(y)
0.99624
```

Ex. 4: model selection by AIC

- Which of the following model fits best to the data in the table ? Answer the question using AIC. Next slide provides hints.
  - Linear model $y=\beta_0+\beta_1 x$
  - Quadratic model $y=\beta_0+\beta_1 x+\beta_2 x^2$
  - Cubic model $y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3$
  - 4th polynomial $y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^4$
  - 5th polynomial $y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^4+\beta_5 x^5$

x	600	700	800	950	1100	1300	1500
y	253	337	395	451	495	534	573

How to build higher polynomial models

- We can build higher polynomial models by the same strategy as we added a bias term, namely, by modifying the design matrix.

$$\begin{pmatrix} 1\, x_1\, x_1^2 \\ 1\, x_2\, x_2^2 \\ 1\, x_3\, x_3^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 \\ \beta_0 + \beta_1 x_3 + \beta_2 x_3^2 \end{pmatrix}$$

- Example in octave (watch out spaces !)
  - . before ^ indicates that it is an operation to the elements in the matrix. So A.^2 and A^2 differs.

```
X = [1 2 3]'
X = [ones(3,1) X X.^2 X.^3]
```

Appendix

Bias-variance decomposition

~theoretical justification for regularization~

Bias-Variance Decomposition

Target label	$y_i = f(x_i) + \epsilon$
Noise in observation	ϵ
Prediction	$\hat{y}_i$

$$E[MSE] = E[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[(y_i - f(x_i))^2] + E[(E[\hat{y}_i] - \hat{y}_i)^2] + E[(E[\hat{y}_i] - f(x_i))^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[\epsilon^2] + Variance + Bias^2$$

- We assume that target label is observed with some measurement noise ε (irreducible error).
- Note that bias and variance cannot be minimized simultaneously, hence model selection is necessary.
- Details of derivation is shown in the next slide.

$$E[MSE]=E[\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2]=\frac{1}{n}\sum_{i=1}^{n}E[(y_i-\hat{y}_i)^2]$$

$$E[(y_i-\hat{y}_i)^2]=E[(y_i-f(x_i)+f(x_i)-\hat{y}_i)^2]$$

$$=E[(y_i-f(x_i))^2]+E[(\hat{y}_i-f(x_i))^2]+E[(y_i-f(x_i))(f(x_i)-\hat{y}_i)]$$

$$=E[\epsilon^2]+\underline{E[(\hat{y}_i-f(x_i))^2]}$$

$$\mathit{since}\,E[(y_i-f(x_i))(f(x_i)-\hat{y}_i)]$$

$$=E[y_i f(x_i)]-E[y_i\hat{y}_i]-E[f(x_i)^2]+E[f(x_i)\hat{y}_i]$$

$$=E[f(x_i)^2]-E[\hat{y}_i f(x_i)]-E[f(x_i)^2]-E[\hat{y}_i f(x_i)]=0$$

$$\underline{E[(\hat{y}_i-f(x_i))^2]}=E[(\hat{y}_i-E[\hat{y}_i]+E[\hat{y}_i]-f(x_i))^2]$$

$$=E[(\hat{y}_i-E[\hat{y}_i])^2]+E[(E[\hat{y}_i]-f(x_i))^2]+\underline{E[(\hat{y}_i-E[\hat{y}_i])(E[\hat{y}_i]-f(x_i))]}$$

$$=E[(\hat{y}_i-E[\hat{y}_i])^2]+E[(E[\hat{y}_i]-f(x_i))^2]$$

$$=\mathit{Variance}(\hat{y})+\mathit{Bias}(f(x))^2$$

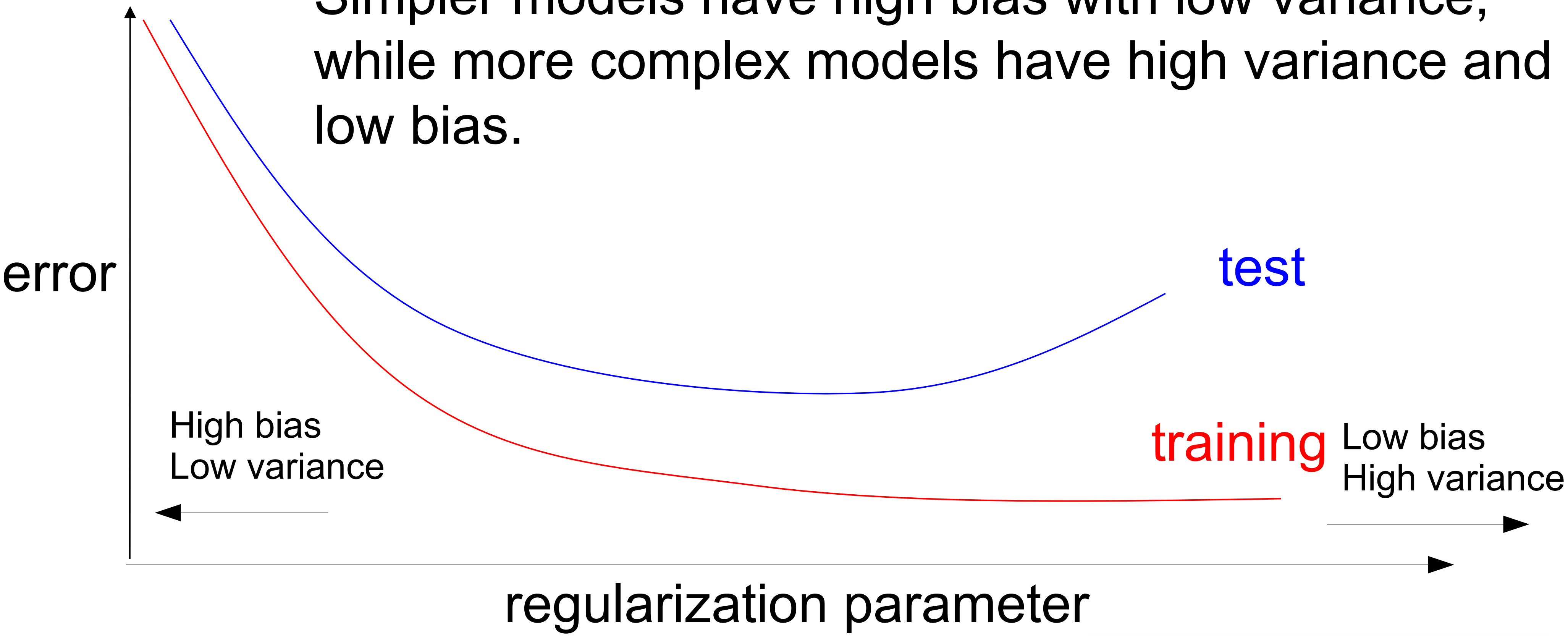
$$\mathit{since}\,\underline{E[(\hat{y}_i-E[\hat{y}_i])(E[\hat{y}_i]-f(x_i))]}$$

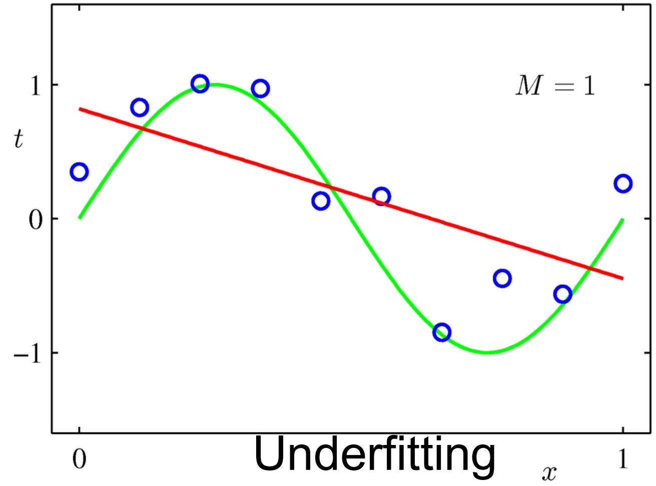
$$=E[\hat{y}_i E[\hat{y}_i]]-E[E[\hat{y}_i]^2]-E[\hat{y}_i f(x_i)]+E[f(x_i)E[\hat{y}_i]]$$

$$=E[\hat{y}_i]^2-E[\hat{y}_i]^2-f(x_i)E[\hat{y}_i]+f(x_i)E[\hat{y}_i]=0$$

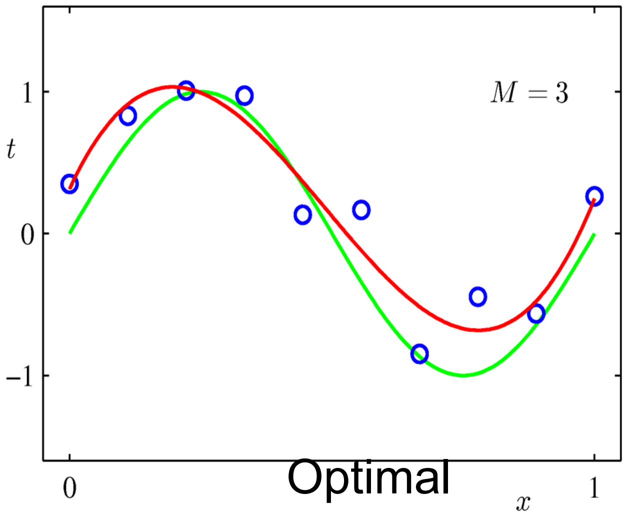
sketch of bias variance trade-off

Simpler models have high bias with low variance, while more complex models have high variance and low bias.

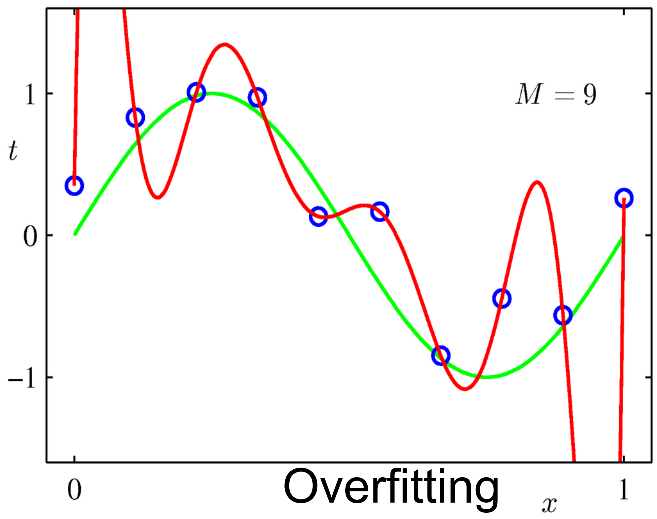




Underfitting



Optimal



Overfitting

Appendix

aproximmate lasso solution by iteratively reweighted ridge regression

Solving LASSO $min_{\beta} \lambda \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$

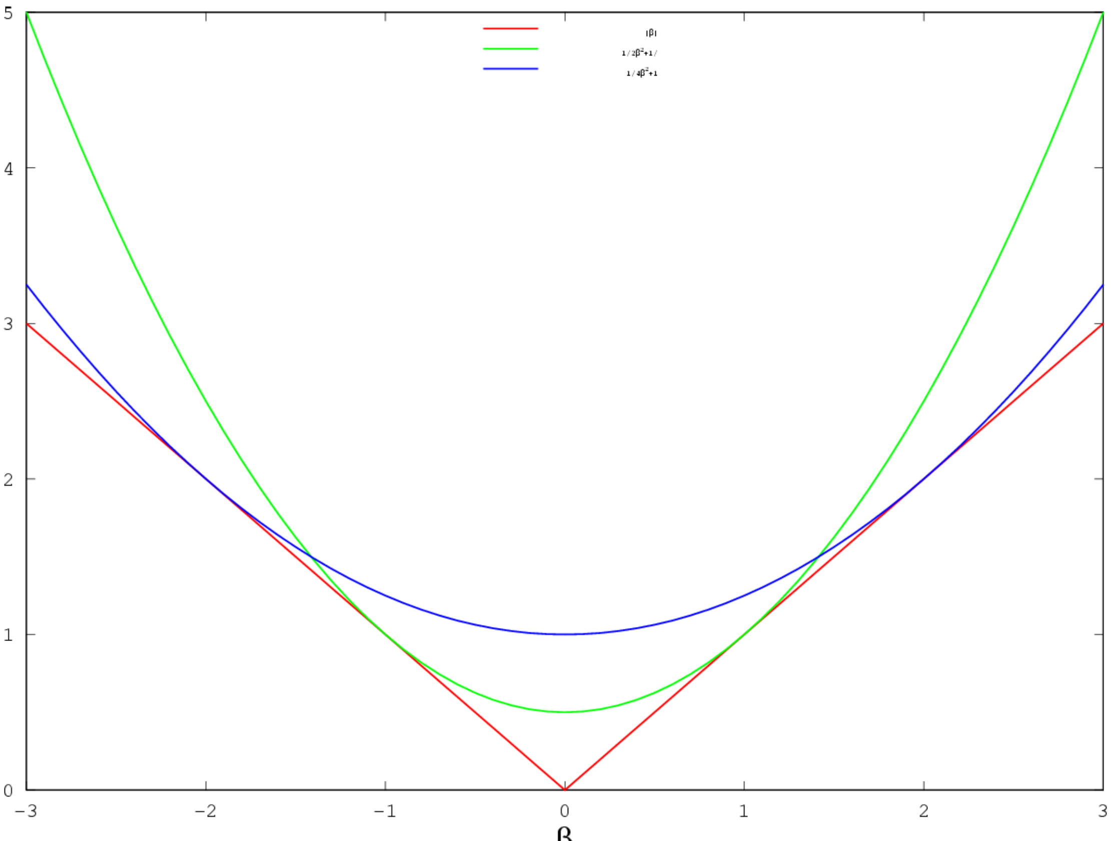
$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

Remember that we have an absolute function in the objective.
It is hard to minimize directly, so we instead minimize its upperbound.

$$\|\boldsymbol{\beta}\|_1 \leq \frac{\beta^2}{2c} + \frac{c}{2}$$

$$c \geq 0$$

Figure shows that
Absolute function (red) is upperbounded by
its approximations
(c=1: green c=2: blue)



Solving LASSO $min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$

$$\|\boldsymbol{\beta}\|_1 \leq \frac{\beta^2}{2c} + \frac{c}{2,} c \geq 0$$

- We now approximate positive constant c by our current solution, then

$$\|\boldsymbol{\beta}\|_1 \leq \frac{\beta^2}{2|\hat{\beta}|} + \frac{|\hat{\beta}|}{2}$$

New objective function is obtained as

$$min_{\boldsymbol{\beta}} \lambda \boldsymbol{\beta} \hat{\mathbf{B}}^{-1} \boldsymbol{\beta} + \frac{\hat{\boldsymbol{\beta}}}{2} + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$$

where $\hat{\mathbf{B}}^{-1}$ is a digonal matrix containing $1/|\hat{\boldsymbol{\beta}}|$ in its diagonal, and the 2nd term can be ignored since β is not a function of $\hat{\boldsymbol{\beta}}$.

Solving LASSO $min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$

$$min_{\boldsymbol{\beta}} \lambda \boldsymbol{\beta} \hat{\mathbf{B}}^{-1} \boldsymbol{\beta} + \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2$$

- An update rule to solve the above optimization problem is obtained as

$$\boldsymbol{\beta}^{lasso} \leftarrow \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \mathbf{B}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

where $\mathbf{B} = diag\left(1/|\beta_1|, 1/|\beta_2|, \ldots, 1/|\beta_p|\right)$

- One can set the initial parameters to any value, for example, $\boldsymbol{\beta} = \mathbf{0}$