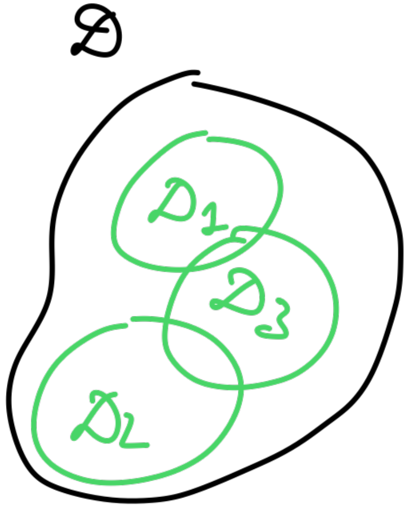
Today

- → illustration of overfitting
- → Bias Variance decomposition
- → regularization
  - → Best subset selection (cross validation)
  - → Ridge regression
  - → LASSO
- → + statistical intuition (Ridge/LASSO)

Bias Variance Decomposition

$h_\beta(x; \mathcal{D}_i)$: hypothesis / model learned on the training set $\mathcal{D}_i \subseteq \mathcal{D} = \{(x^{(i)}, t^{(i)})\}$

To measure how good a hypothesis $h_\beta(x)$ is for a dataset $\mathcal{D}$, we consider MSE for a new x



$$MSE(x) = \mathbb{E}_{\mathcal{D}_i}\left\{\left(t(x) - h_\beta(x; \mathcal{D}_i)\right)^2\right\}$$

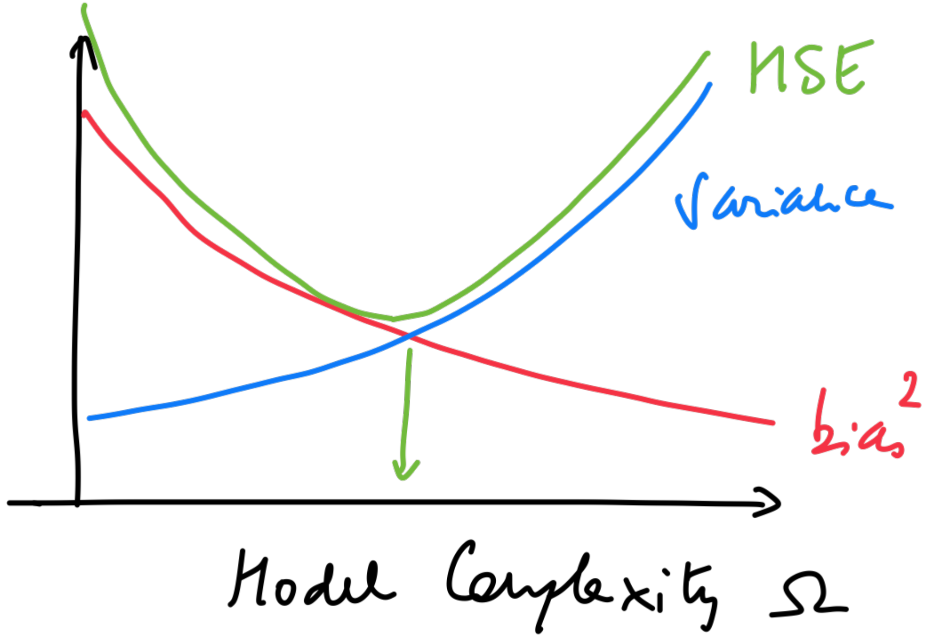
$$MSE(x) = \mathbb{E}_{\mathcal{D}_i}\Big\{\big(t(x) - \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i) + \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i) - h_\beta(x;\mathcal{D}_i)\big)^2$$

$$= \mathbb{E}_{\mathcal{D}_i}\Big\{\big(t(x) - \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i)\big)^2\Big\} + \mathbb{E}_{\mathcal{D}_i}\Big\{\big(\mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i) - h_\beta(x;\mathcal{D}_i)\big)^2$$

$$+ 2\,\mathbb{E}_{\mathcal{D}_i}\Big\{\big(t(x) - \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i)\big)\big(\mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i) - h_\beta(x;\mathcal{D}_i)\big)\Big\}$$

$$= 0$$

$$= \underbrace{\big(t(x) - \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i)\big)^2}_{\text{bias}^2} + \mathbb{E}_{\mathcal{D}_i}\Big\{\underbrace{\big(h_\beta(x;\mathcal{D}_i) - \mathbb{E}_{\mathcal{D}_i} h_\beta(x;\mathcal{D}_i)\big)^2}_{\text{variance}}\Big\}$$



How can we automatically select the optimal complexity?

→ Regularization (3 most popular approaches)

→ Best Subset Selection
→ Ridge regression
→ LASSO

→ Main objective: control the variance while keeping a relatively low bias

Best Subset Selection

⊖ intractable / Combinatorial for large D

Find the optimal subset of the features

→ Total number $\binom{D}{K}$ for every $K = 1, \dots, D$

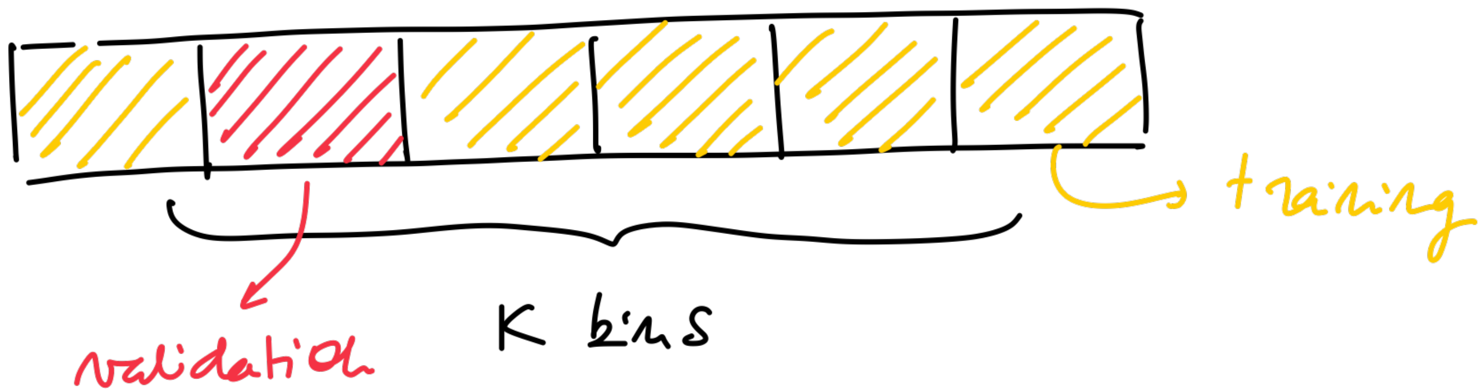
For each subset → ① fit the model to the dataset

② Evaluate trained model on some new (test) dataset

③ Select the subset of the features that gives the best prediction error.

→ Can be implemented through cross validation
(K fold cross validation)

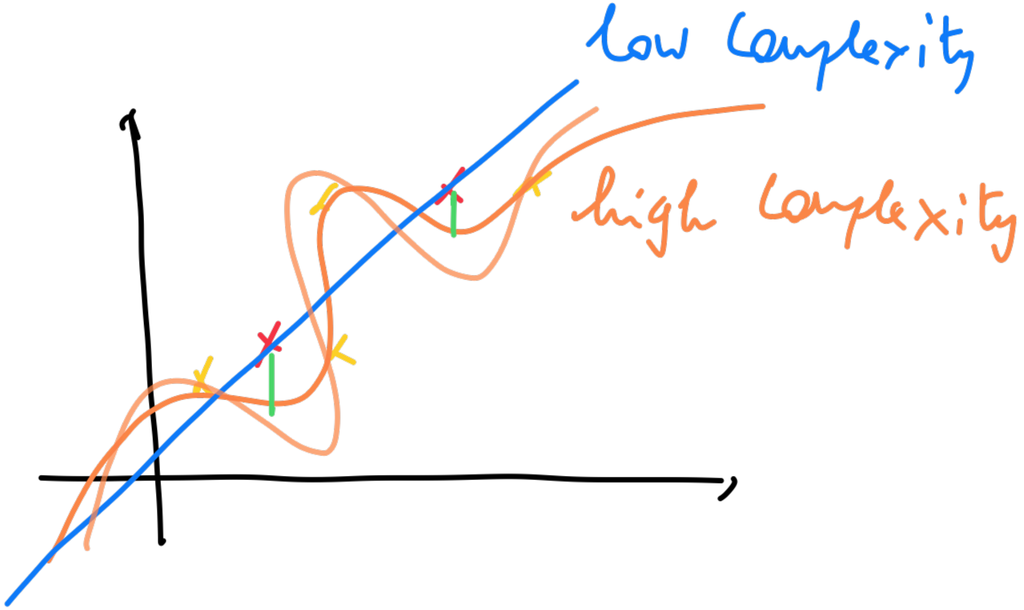
→ <u>Step 1</u> : Split the data set into K bins



For each bin $k = 1, \dots, K$ train the model on all the bins but the k^{th} one and evaluate it on the k^{th} bin

Leave-one-out cross validation : $K = 1$ / train the model on $\mathcal{D}$ but one example and evaluate hp on the remaining example.

$$\text{error}_{CV} = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_{\beta}^{-k(i)} \left(x^{(i)} \right) \right)^2$$



→ 2 Alternatives : Ridge , LASSO

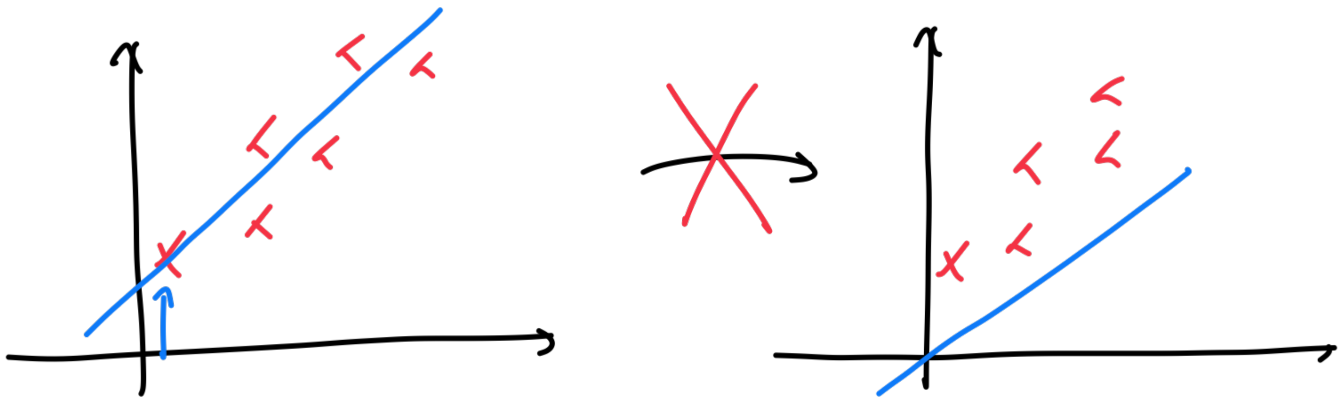
following from addition of a penalty to the OLS loss

Recall $\ell_{OLS}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)}) \right)^2$

$= \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \beta^T \tilde{x}^{(i)} \right)^2$

<u>Ridge loss</u> : $\ell_{Ridge}(\beta) = \overbrace{\frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \beta^T \tilde{x}^{(i)} \right)^2}^{\text{data fidelity}} + \underbrace{\lambda \sum_{j=1}^{D} |\beta_j|^2}_{\text{Penalty on Model Complexity}}$

⚠ We do not penalize the intercept β_0



LASSO : $\ell_{LASSO}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \beta^T \tilde{x}^{(i)}\right)^2 + \underbrace{\lambda \sum_{j=1}^{D} |\beta_j|}$

→ Complexity: Ridge can be solved through gradient descent (differentiable everywhere)

In fact we can get the β_{Ridge} (regression vector that minimizes the Ridge loss) through the resolution of a linear system

Developing the Ridge loss as we did it for the OLS

$$\ell_{Ridge}(\beta) = \frac{1}{N} e^T e + \lambda \sum_{j=1}^{D} |\beta_j|^2$$

$$= \frac{1}{N} \underbrace{(t - \tilde{\underline{\underline{X}}}\beta)^T (t - \tilde{X}\beta)} + \lambda \underbrace{\sum_{j=1}^{D} |\beta_j|^2}$$

→ First solving for the intercept, we get

$$\frac{\partial}{\partial \beta_0} \frac{1}{N} \sum \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2$$

$$\frac{1}{N} \sum \left(t^{(i)} - (\beta_0 + \beta_1 x^{(i)} + \dots + \beta_D x_D^{(i)})\right)(-2) = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} t^{(i)} = \beta_0 + \sum_{i=1}^{N}\sum_{j=1}^{D} x_j^{(i)}\beta_j$$

if $x^{(i)}$ are centered, $\sum_{i=1}^{N} x_j^{(i)} = 0$

then $\beta_0 = \frac{1}{N}\sum_{i=1}^{N} t^{(i)}$

For centered $x^{(i)}$'s, the Ridge loss can read as

$$\ell_{Ridge}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \frac{1}{N}\sum_{i=1}^{N} t^{(i)} - \beta_{1\to D}^{T} x^{(i)}\right)^2 + \lambda\sum_{j=1}^{D}|\beta_j|^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\underbrace{t^{(i)} - \bar{t}}_{\tilde{t}^{(i)}} - \beta_{1\to D}^{T} x^{(i)}\Big)^2 + \lambda\sum_{j=1}^{D}|\beta_j|^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{t}^{(i)} - \beta_{1\to D}^{T}\, x^{(i)}\right)^2 + \lambda \,\|\beta_{1\to D}\|_2^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{t}^{(i)} - \beta_{1\to D}^{T}\, x^{(i)}\right)^2 + \lambda \,(\beta_{1\to D})^T(\beta_{1\to D})$$

→ As for the OLS loss, we can find β_{Ridge} directly by computing grad ℓ_{Ridge} and set it to 0.

$$\ell_{Ridge} = \frac{1}{N}\left(\tilde{t} - \underline{\underline{X}}\,\beta_{1\to D}\right)^T\left(\tilde{t} - \underline{\underline{X}}\,\beta_{1\to D}\right) + \lambda\,\underbrace{(\beta_{1\to D})^T(\beta_{1\to D})}$$

$$= \frac{1}{N}\,\tilde{t}^T\tilde{t} - 2\,\overbrace{\tilde{t}^T\underline{\underline{X}}\,\beta_{1\to D}} + \beta_{1\to D}^T\,\underline{\underline{X}}^T\underline{\underline{X}}\,\beta_{1\to D}$$

$$+ \lambda\,\beta_{1\to D}^T\,\beta_{1\to D}$$

$$\underset{\beta}{\text{grad}}\, \ell_{Ridge} = -2\underline{\underline{X}}^T\tilde{t} + 2\underline{\underline{X}}^T\underline{\underline{X}}\beta_{1\to D} + 2\lambda\beta_{1\to D} = 0$$

$$\Rightarrow 2\underbrace{\left(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda I\right)}\beta_{1\to D} = 2\underline{\underline{X}}^T\tilde{t}$$

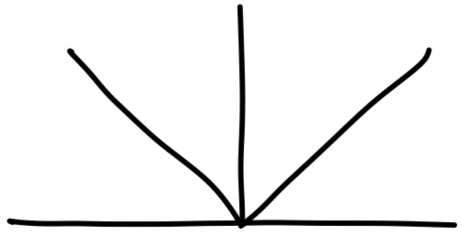
$$\boxed{\beta_{1\to D,\,Ridge} = \left(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda I\right)^{-1}\underline{\underline{X}}^T\tilde{t}}$$

Advantage of Ridge vs OLS: even if $\underline{\underline{X}}^T X$ was not invertible (in OLS because of redundancy in features or high complexity model), as soon as $\lambda > 0$, the matrix $(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda I)$ which shifts the eigenvalues of $\underline{\underline{X}}^T\underline{\underline{X}}$ by $\lambda > 0$ is always invertible.

$$\beta_{Ridge} = \begin{cases} \beta_0 = \frac{1}{N}\sum_{i=1}^{N} t^{(i)} \\ \beta_{1\to D} = \left(\underline{\underline{X}}^T \underline{\underline{X}} + \lambda I\right)^{-1} \underline{\underline{X}}^T \tilde{t} \end{cases}$$

Where $\tilde{t} = t - \frac{1}{N}\sum_{i=1}^{N} t^{(i)}$

For LASSO, note that $|\beta_j|$ is not differentiable at zero



→ gradient descent will not work

→ β_{LASSO} cannot be obtained from solving a linear system unlike OLS and Ridge.

→ However LASSO will be better at performing feature selection.