Ridge Regression

Regularization in Linear Regression

Recall Least Squares:

$$\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

when $(\mathbf{X}^T\mathbf{X})^{-1}$ exists….

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$\hat{w} = \arg\min_w \|w\|_2^2$

s.t. $X^T X w = X^T y$

$X^T X w = X^T y$

Regularization in Linear Regression

Recall Least Squares:

$$\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

In general:

$$= \arg\min_w w^T (\mathbf{X}^T \mathbf{X}) w - 2y^T \mathbf{X} w$$

Regularization in Linear Regression

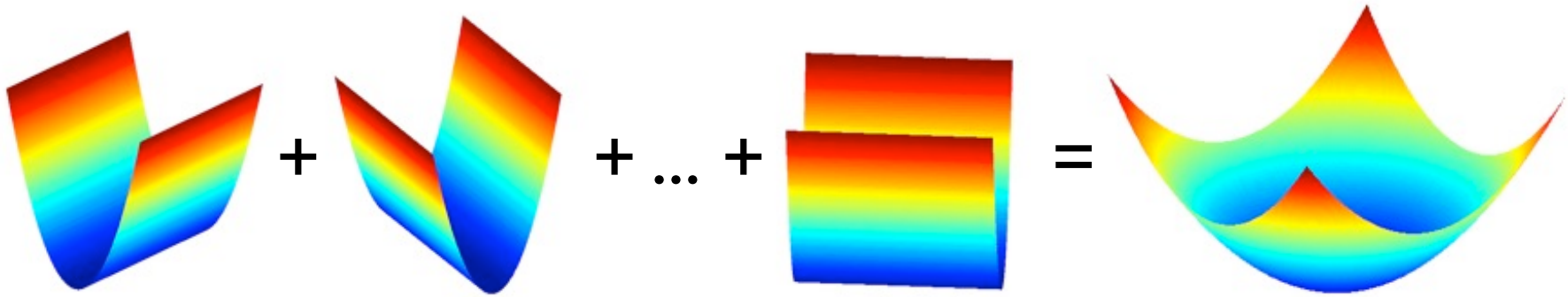
Recall Least Squares:

$$\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

In general:

$$= \arg\min_w w^T (\mathbf{X}^T \mathbf{X}) w - 2y^T \mathbf{X} w$$



$$(y_1 - x_1^T w)^2 + (y_2 - x_2^T w)^2 + \cdots + (y_n - x_n^T w)^2 = \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

$w \in \mathbb{R}^2$

What if $x_i \in \mathbb{R}^d$ and $d > n$?

Regularization in Linear Regression

Recall Least Squares: $$\widehat{w}_{LS} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

When $x_i \in \mathbb{R}^d$ and $d > n$ the objective function is flat in some directions:

Regularization in Linear Regression

Recall Least Squares: $\widehat{w}_{LS} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$

When $x_i \in \mathbb{R}^d$ and $d > n$ the objective function is flat in some directions:

Implies optimal solution is *not unique* and unstable due to lack of curvature*:*

- small changes in training data result in large changes in solution
- often the *magnitudes* of *w* are "very large"

Regularization imposes "simpler" solutions by a "complexity" penalty

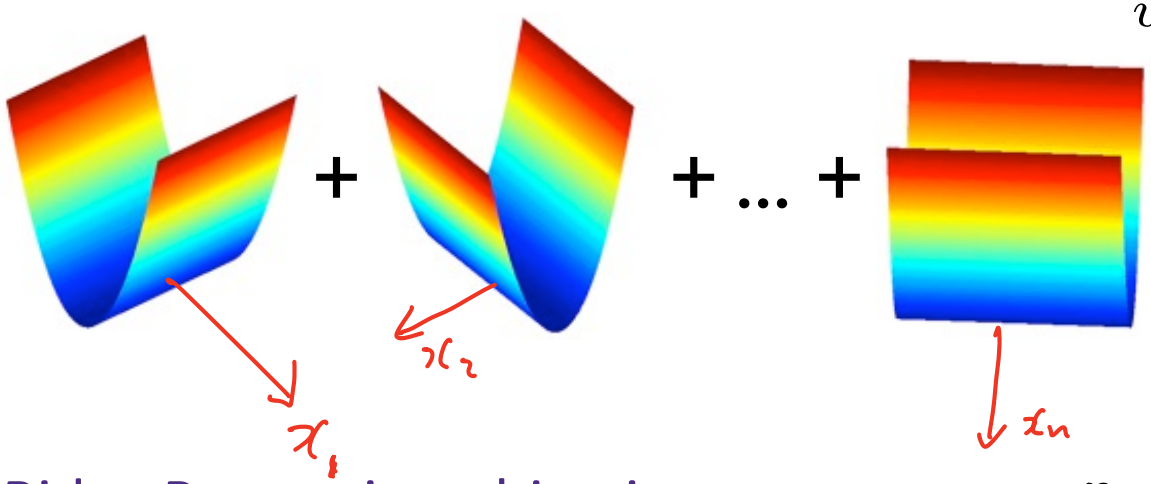
Ridge Regression

$e_i = [0 \; 0 \; \dots \; 1 \; \dots \; 0]$ ← i^{th} position

$x_i = e_i$

- Old Least squares objective:

$$\hat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$



$(e_i^T w - y_i)^2 \quad y_i = 1$

$= w^T e_i e_i^T w - 2 e_i^T w + 1$

$= w_1^2$

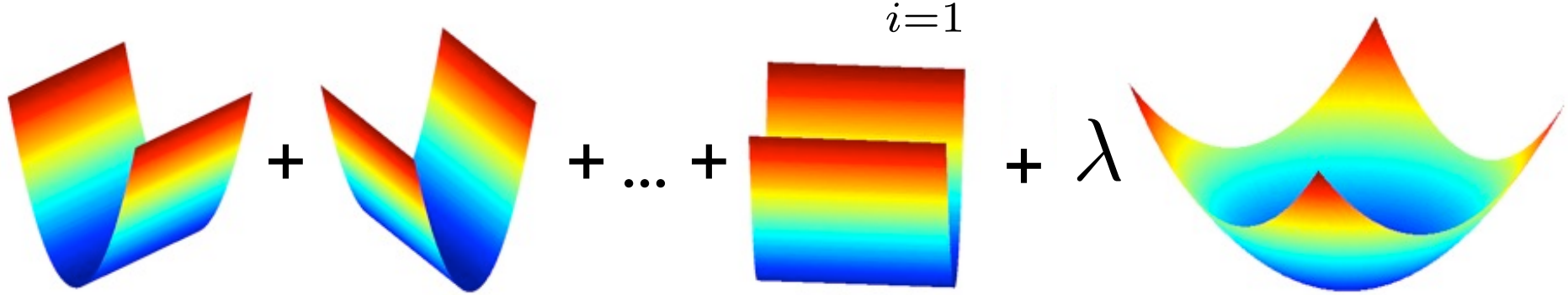
$\sum_{i=1}^{d} (e_i^T w - 1)^2 = \sum_i w^T e_i e_i^T w - 2 e_i^T w + 1$

$= w^T \left(\sum_i e_i e_i^T\right) w - \dots$

$= \|w\|_2^2$

- Ridge Regression objective:

$$\hat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$



Minimizing the Ridge Regression Objective

$$\|w\|_2^2 = \sum_{i=1}^{d} w_i^2 = w^T w = w^T I w$$

$$\hat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

$$= \left(\sum_{i=1}^{n} y_i^2 - 2 x_i^T w y_i + w^T x_i x_i^T w \right) + \lambda w^T w$$

$$= \left(\sum_{i=1}^{n} y_i^2 \right) - 2 \left(\sum_{i=1}^{n} x_i y_i \right)^T w + w^T \left(\sum_{i=1}^{n} x_i x_i^T \right) w + \lambda w^T w$$

$$= \left(\sum_{i=1}^{n} y_i^2 \right) - 2 \left(\sum_{i=1}^{n} x_i y_i \right)^T w + w^T \left(\lambda I + \sum_{i=1}^{n} x_i x_i^T \right) w = (*)$$

$$\nabla_w (*) = -2 \left(\sum_{i=1}^{n} x_i y_i \right) + 2 \left(\lambda I + \sum_{i=1}^{n} x_i x_i^T \right) w = 0$$

$$\hat{w}_{Ridge} = \left(\sum_{i=1}^{n} x_i x_i^T + \lambda I \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right)$$

Shrinkage Properties

$$\widehat{w}_{ridge} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

$$= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

$$= \frac{1}{\lambda}\left(\frac{1}{\lambda}X^TX + I\right)^{-1}X^Ty$$

$$\|Xw - y\|_2^2 + \lambda\|w\|_2^2 = w^TX^TXw - 2w^TX^Ty + y^Ty + \lambda w^Tw$$

$$= w^T\left(X^TX + \lambda I\right)w - 2X^Ty + y^Ty = (*)$$

$$\nabla_w (*) = 2\left(X^TX + \lambda I\right)w - 2X^Ty = 0$$

$$\left(X^TX + \lambda I\right)w = X^Ty$$

Bias-Variance Properties

$x_i \sim \mathcal{N}(0, I)$

$\Rightarrow \mathbb{E}[x_i x_i^T] = I$

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

Bias-Variance Properties

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

$$\mathbb{E}_{Y|X,\mathcal{D}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$$
$$= \underbrace{\mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x]}_{\text{Irreducible Error}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T\widehat{w}_{ridge})^2]}_{\text{Learning Error}}$$

Bias-Variance Properties

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned}
&\mathbb{E}_{Y|X,\mathcal{D}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x] \\
&\quad = \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T\widehat{w}_{ridge})^2] \\
&\quad = \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T\widehat{w}_{ridge})^2] \\
&\quad = \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}] - x^T\widehat{w}_{ridge})^2]
\end{aligned}$$

Irreduc. Error — Bias-squared — Variance

Bias-Variance Properties

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned}
&\mathbb{E}_{Y|X,\mathcal{D}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x] \\
&\quad = \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T\widehat{w}_{ridge})^2] \\
&\quad = \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T\widehat{w}_{ridge})^2] \\
&\quad = \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}] - x^T\widehat{w}_{ridge})^2]
\end{aligned}$$

Irreduc. Error — Bias-squared — Variance

$$\begin{aligned}
\widehat{w}_{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T(\mathbf{X}w + \boldsymbol{\epsilon}) \\
&= \frac{n}{n+\lambda}w + \frac{1}{n+\lambda}\mathbf{X}^T\boldsymbol{\epsilon}
\end{aligned}$$

Bias-Variance Properties

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned}
&\mathbb{E}_{Y|X,\mathcal{D}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x] \\
&\quad = \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T\widehat{w}_{ridge})^2] \\
&\quad = \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T\widehat{w}_{ridge})^2] \\
&\quad = \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T\widehat{w}_{ridge}] - x^T\widehat{w}_{ridge})^2] \\
&\quad = \sigma^2 + \frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2 + \frac{d\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2
\end{aligned}$$

(verify at home)

Irreduc. Error — Bias-squared — Variance

Bias-Variance Properties

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- ***Assume:*** $\mathbf{X}^T\mathbf{X} = nI$ ***and*** $\mathbf{y} = \mathbf{X}w + \boldsymbol{\epsilon}$ $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$

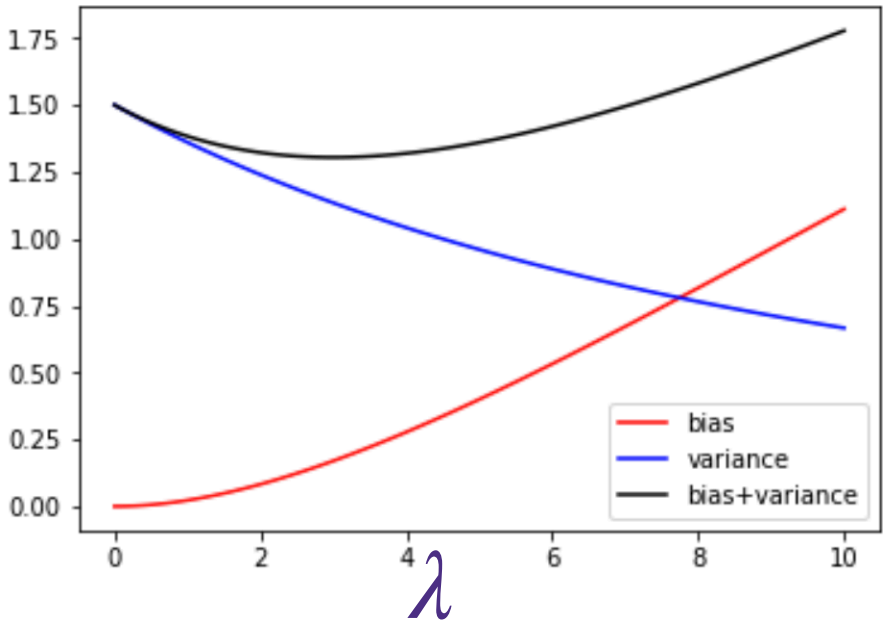
If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x,\text{train}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$?

$$\mathbb{E}_{Y|X,\mathcal{D}}[(Y - x^T\widehat{w}_{ridge})^2 | X = x]$$

$$= \sigma^2 + \frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2 + \frac{d\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2$$

Irreduc. Error — Bias-squared — Variance

(verify at home)



d=10, n=20,
$\sigma^2 = 3.0, \|w\|_2^2 = 10$

Ridge Regression: Effect of Regularizatio

$$\widehat{w}_{ridge} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

- Solution is indexed by the regularization parameter λ
- Larger λ

- Smaller λ

- As λ → 0, $\widehat{w}_{ridge} \rightarrow \hat{w}_{LS}$

- As λ →∞, $\widehat{w}_{ridge} \rightarrow 0$

Ridge Regression: Effect of Regularization

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\widehat{w}^{(\lambda)}_{\mathcal{D},ridge} = \arg\min_{w} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T w)^2 + \lambda ||w||_2^2$$

TRAIN error:

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2$$

TRUE error:

$$\mathbb{E}[(Y - X^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

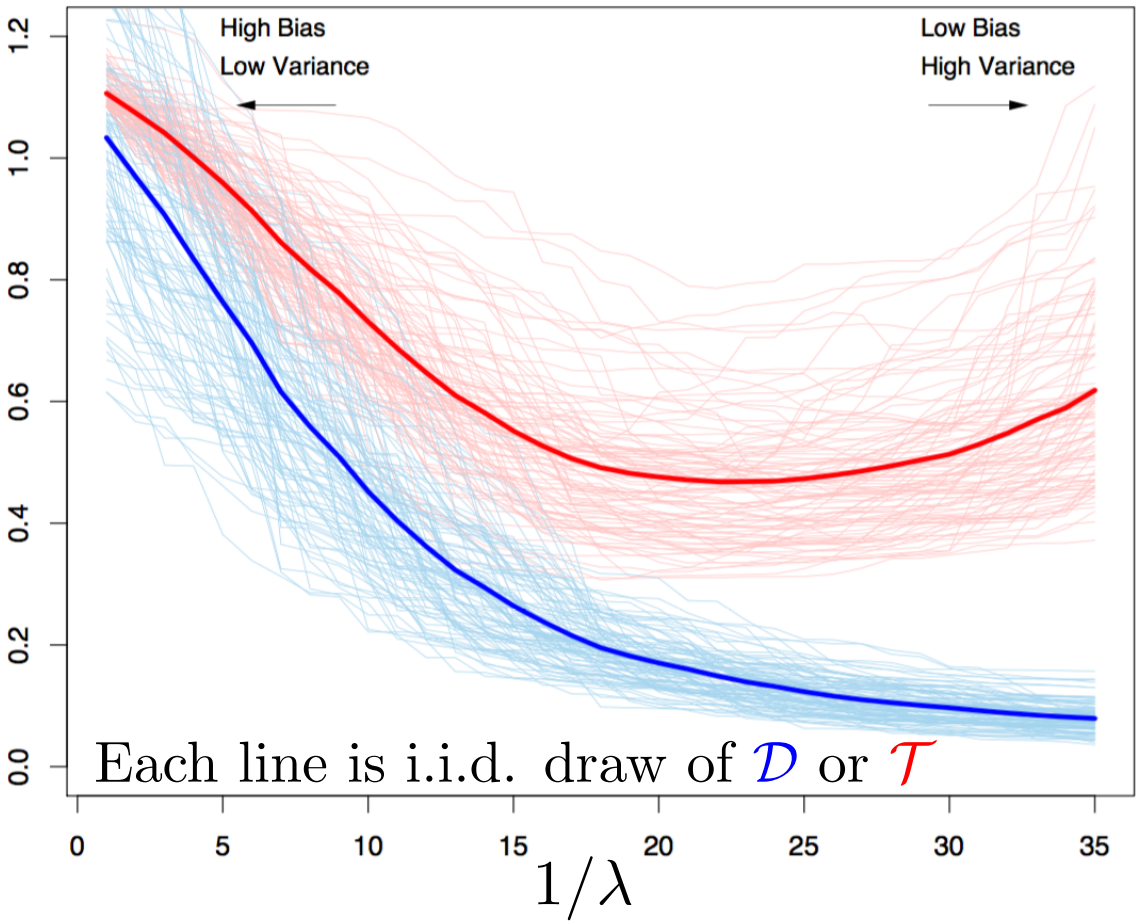
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2$$

Important: $\mathcal{D} \cap \mathcal{T} = \emptyset$

Ridge Regression: Effect of Regularization

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\widehat{w}^{(\lambda)}_{\mathcal{D},ridge} = \arg\min_w \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T w)^2 + \lambda ||w||_2^2$$



TRAIN error:

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2$$

TRUE error:

$$\mathbb{E}[(Y - X^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

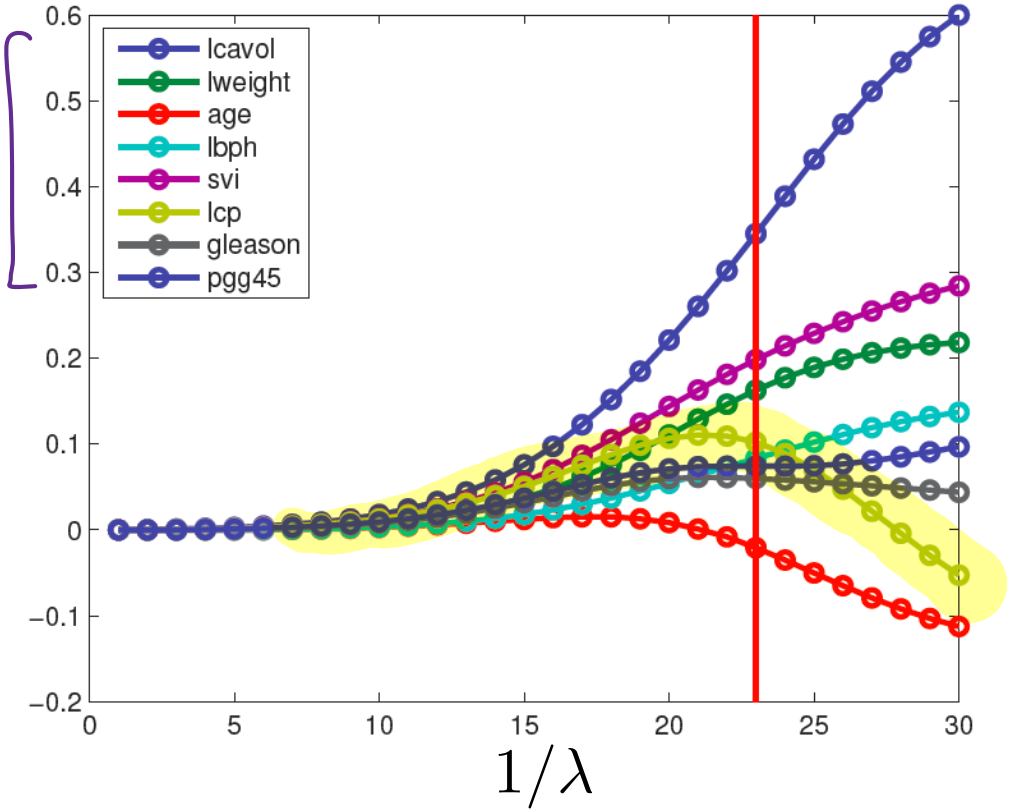
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - x_i^T \widehat{w}^{(\lambda)}_{\mathcal{D},ridge})^2$$

Important: $\mathcal{D} \cap \mathcal{T} = \emptyset$

Ridge regression: $\text{minimize} \sum_{i=1}^{n} (w^T x_i + b - y_i)^2 + \lambda \|w\|_2^2$

$\mathbb{R}^8$

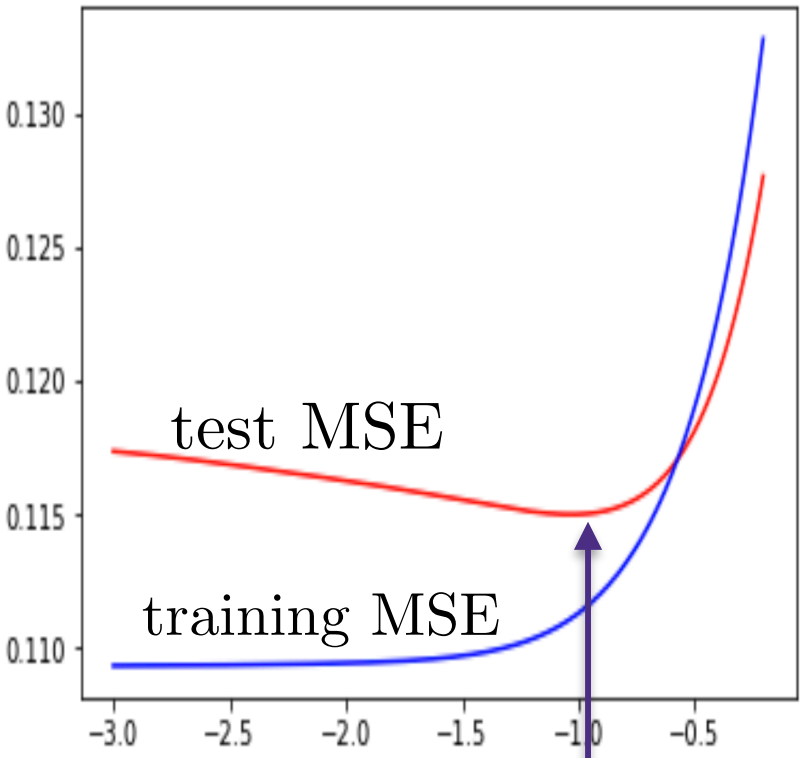
w

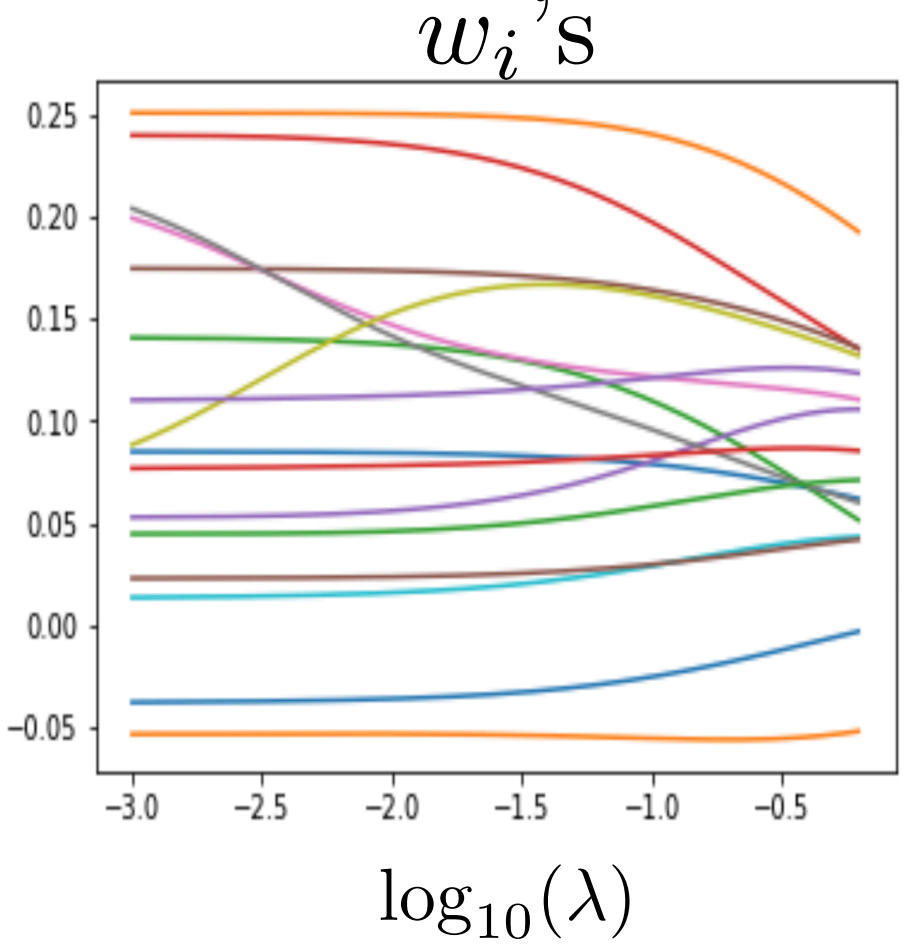


From Kevin Murphy textbook

> **Typical approach: select λ using cross validation, up next**

Ridge regression: minimize $\sum_{i=1}^{n}(w^T x_i + b - y_i)^2 + \lambda\|w\|_2^2$





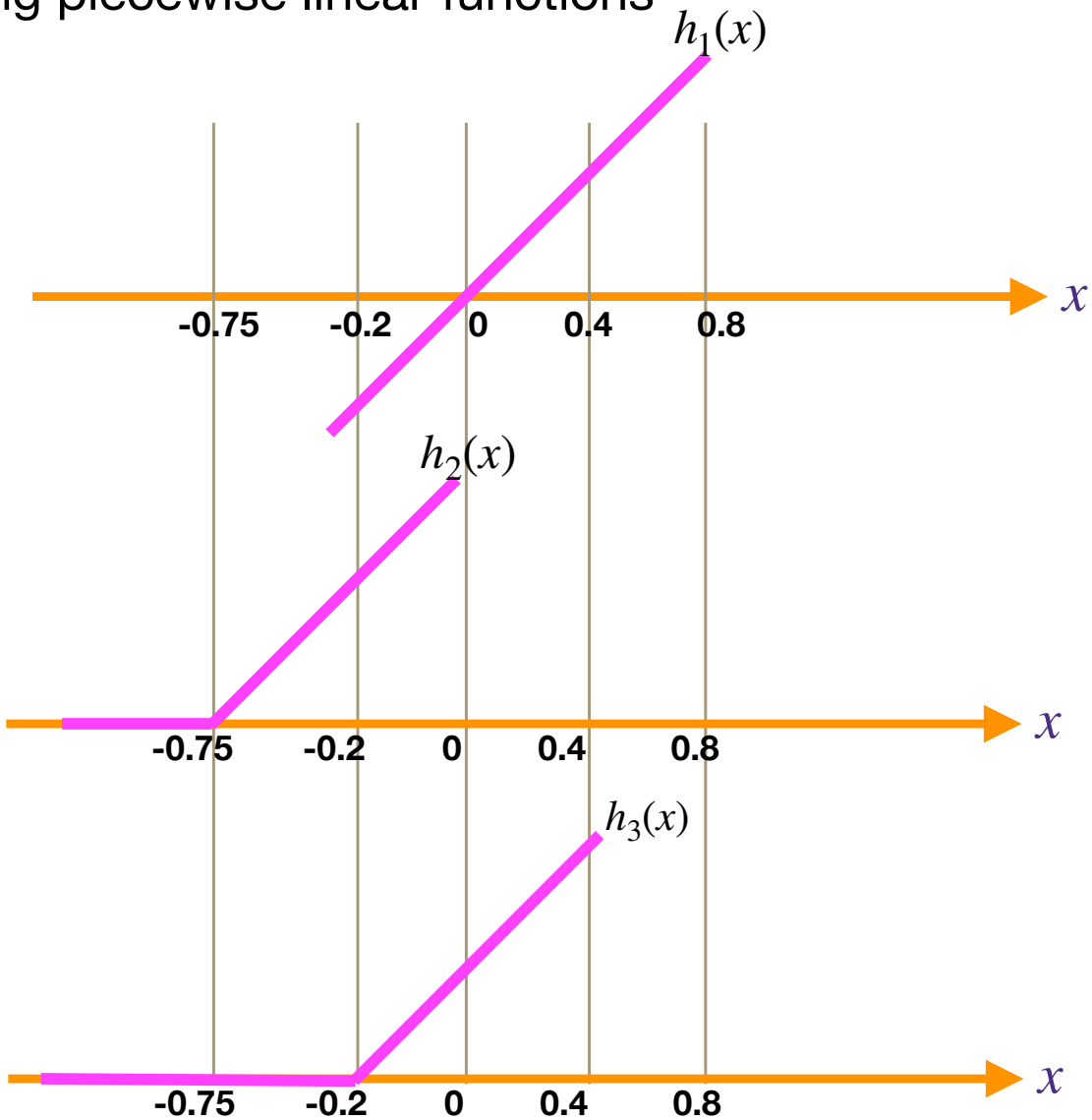
- this gain in test MSE comes from shrinking w's to get a less sensitive predictor
  (which in turn reduces the variance)

Example: piecewise linear fit

- we fit a linear model for $x \in [-1,1]$:
  $f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$
- with a specific choice of features using piecewise linear functions

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ h_3(x) \\ h_4(x) \\ h_5(x) \end{bmatrix} = \begin{bmatrix} x \\ [x+0.75]^+ \\ [x+0.2]^+ \\ [x-0.4]^+ \\ [x-0.8]^+ \end{bmatrix}$$

$$[a]^+ \triangleq \max\{a, 0\}$$

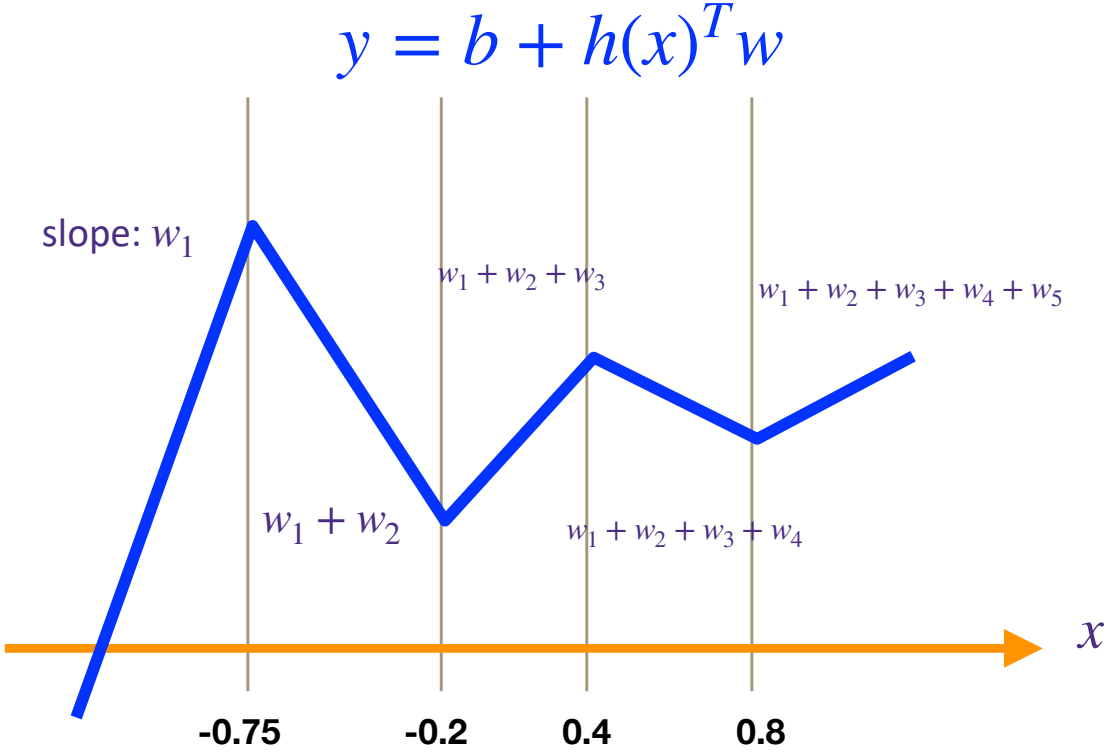


Example: piecewise linear fit

- we fit a linear model:
  $f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$
- with a specific choice of features using piecewise linear functions

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ h_3(x) \\ h_4(x) \\ h_5(x) \end{bmatrix} = \begin{bmatrix} x \\ [x + 0.75]^+ \\ [x + 0.2]^+ \\ [x - 0.4]^+ \\ [x - 0.8]^+ \end{bmatrix}$$

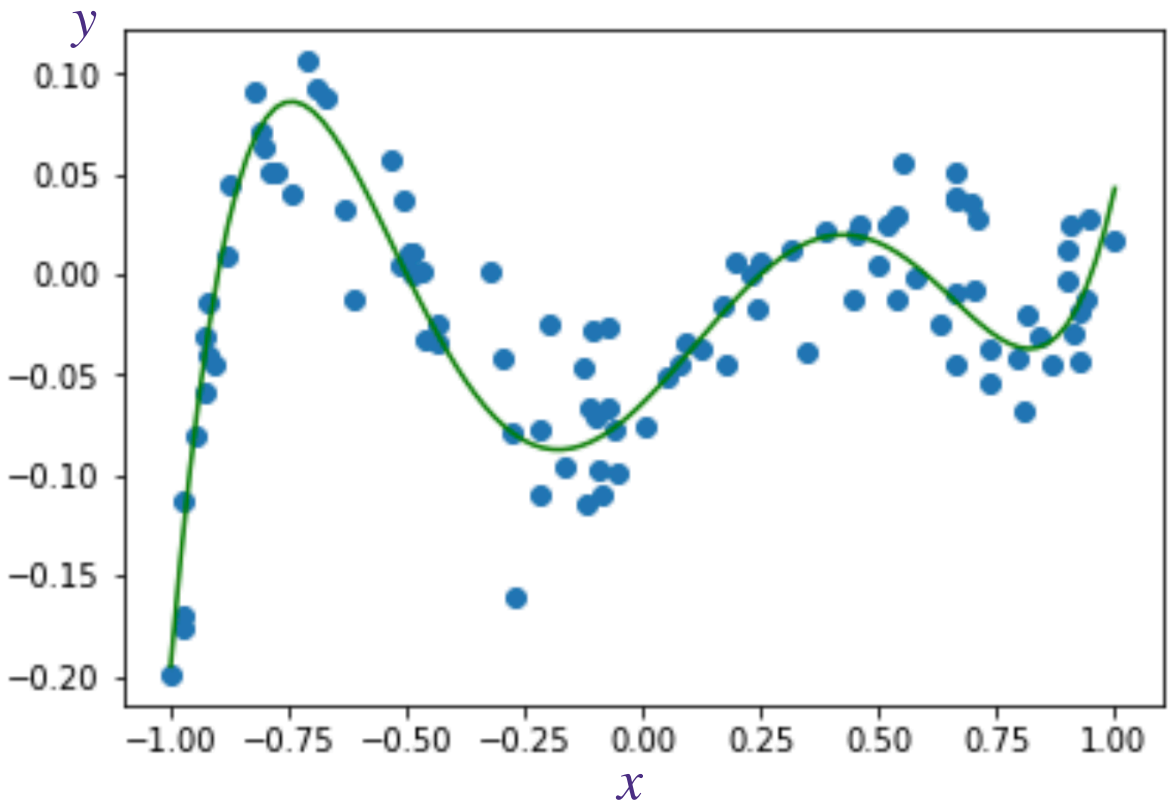
$$[a]^+ \triangleq \max\{a, 0\}$$



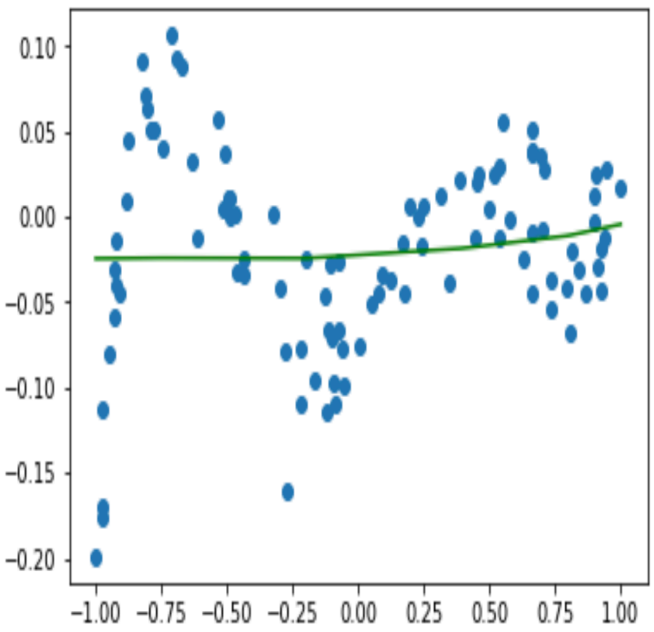
the weights capture the change in the slopes

Example: piecewise linear fit

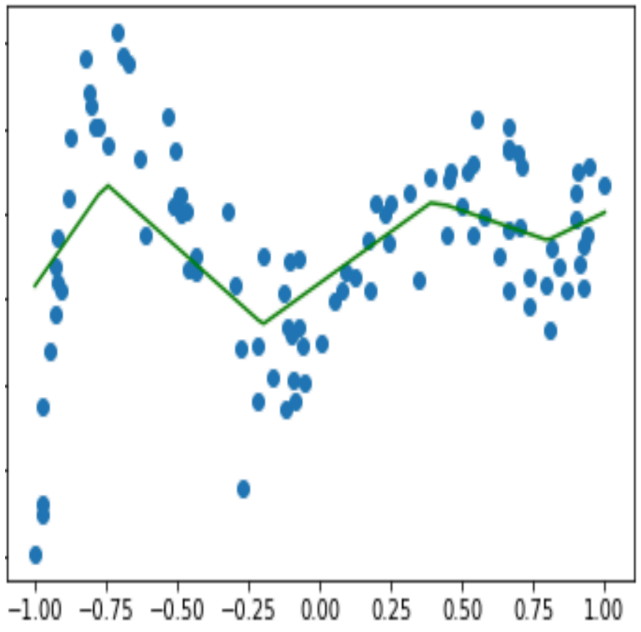
- we fit a linear model:
  $f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$
- with a specific choice of features using piecewise linear functions



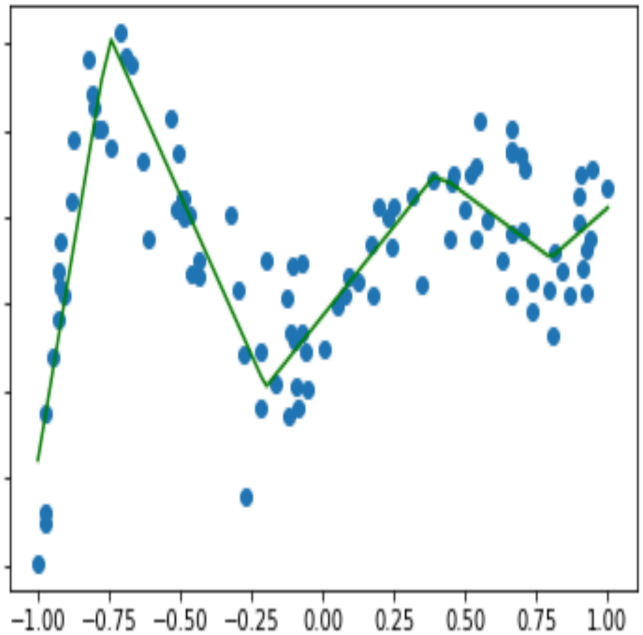
Example: piecewise linear fit (ridge regression)



$\lambda = 1$



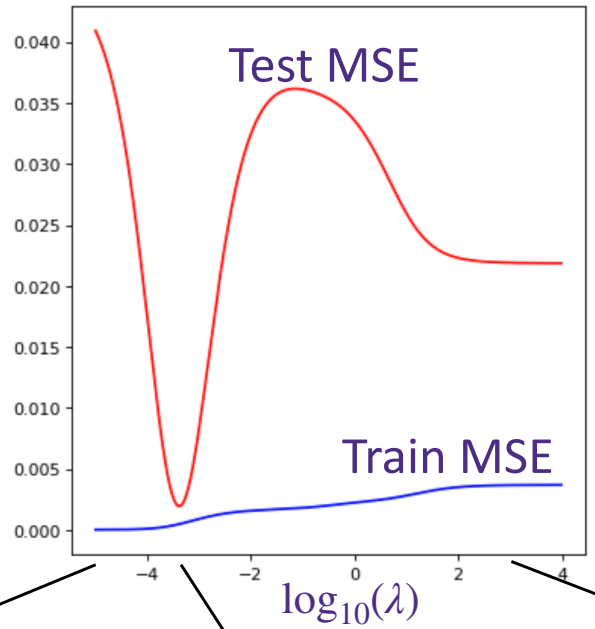
$\lambda = 0.005$

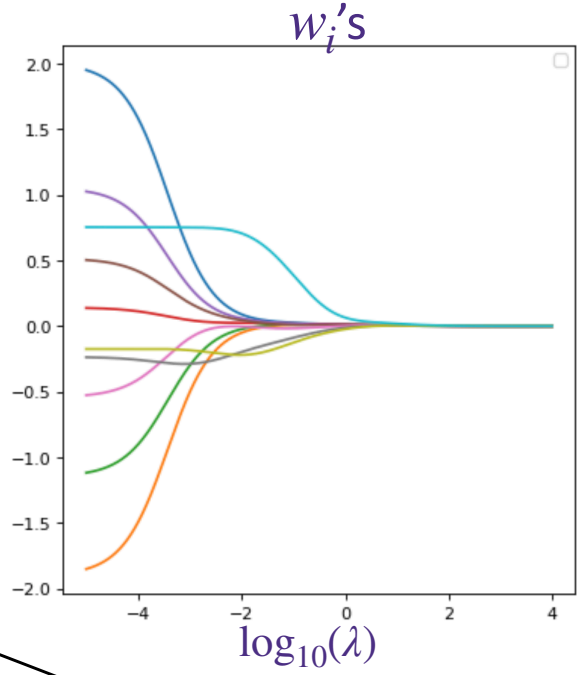


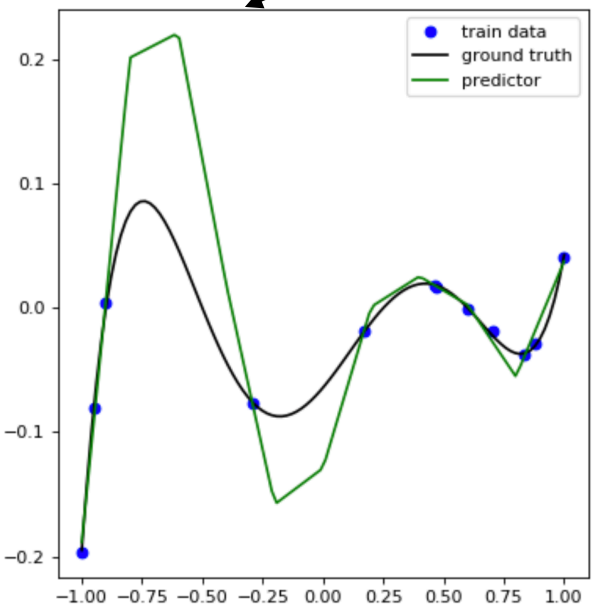
$\lambda = 0.000001$

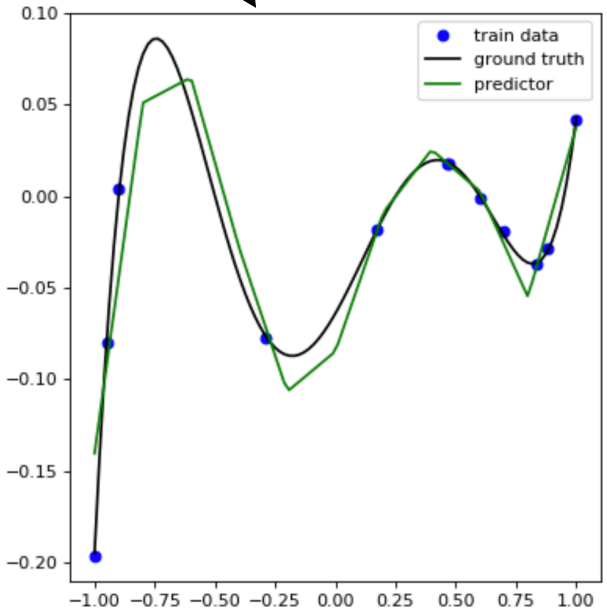
We do not observe overfitting, as d=5 and n=100

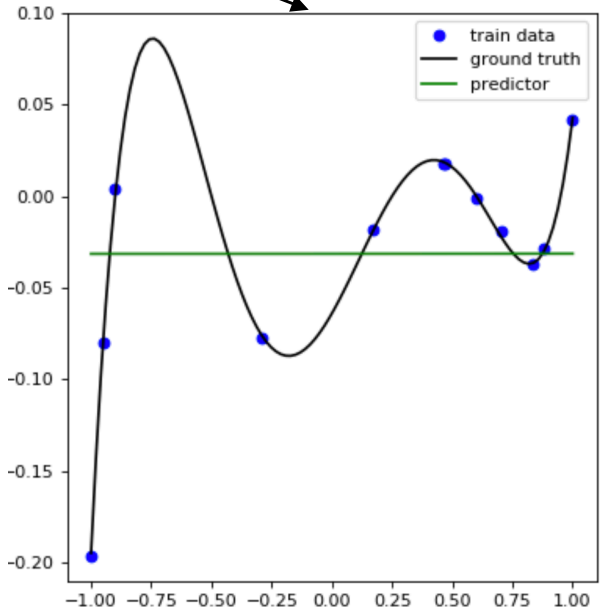
Can avoid overfitting even $w \in \mathbb{R}^{10}$ and n=11 samples











What you need to know…

- Regularization
  - Penalizes complex models towards preferred, simpler models
- Ridge regression
  - L_2 penalized least-squares regression
  - Regularization parameter trades off model complexity with training error
  - Never regularize the offset!