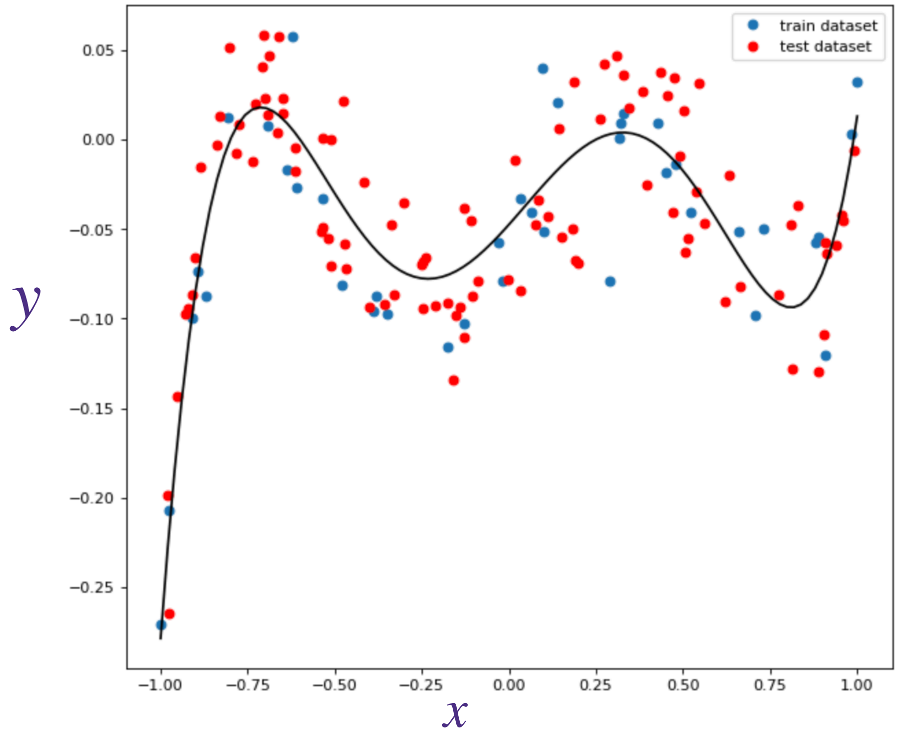
Regularization

Recap: bias-variance tradeoff

- Consider ~~100~~ 40 training examples and 100 test examples i.i.d.drawn from degree-5 polynomial features
$x_i \sim \text{Uniform}[-1,1]$, $y_i \sim f_{w*}(x_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$f_w(x_i) = b^* + w_1^* x_i + w_2^*(x_i)^2 + w_3^*(x_i)^3 + w_4^*(x_i)^4 + w_5^*(x_i)^5$$



This is a linear model with features
$h(x_i) = (x_i, (x_i)^2, (x_i)^3, (x_i)^4, (x_i)^5)$

Choose K-degree.

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

least-squares fit $\rightarrow f_{\hat{w}_{LS}}(x_i, x_i^2, x_i^3)$

$\begin{cases} K=3 \\ K=20 \end{cases}$ $K^* = 5$

Recap: bias-variance tradeoff

$\min_{w,b} \sum_i (Y_i - h(x_i)^T \cdot w - b)^2$

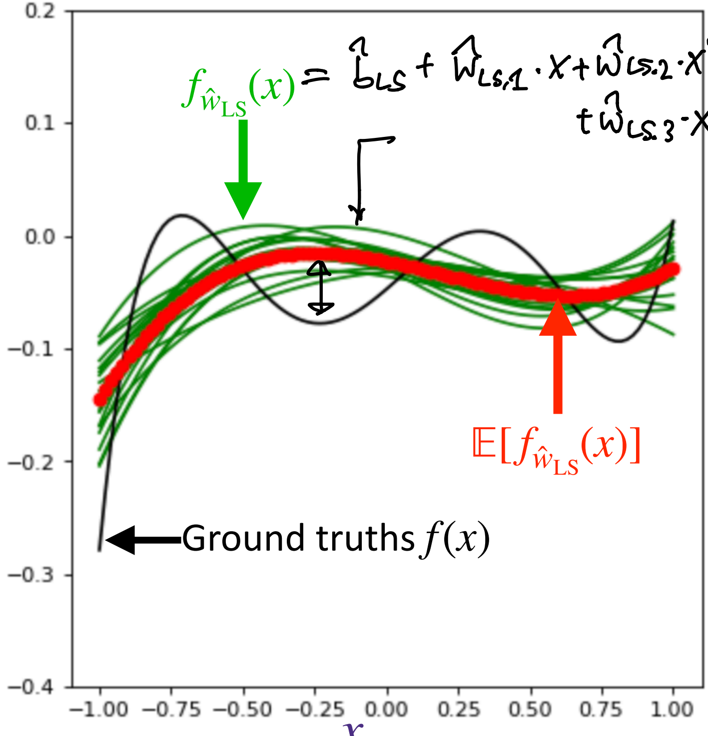
$h(x_i) = (x_i, x_i^2, x_i^3) \in \mathbb{R}^3$, $w \in \mathbb{R}^3$, $b \in \mathbb{R}$. $\to \hat{w}_{LS}, \hat{b}_{LS}$.

With degree-3 polynomials, we underfit

$|\text{Red} - \text{black}| = \text{bias} \uparrow$

$\Delta \text{Green} = \text{variance} \downarrow$

$\hat{f}_{\hat{w}_{\mathrm{LS}}}(x)$



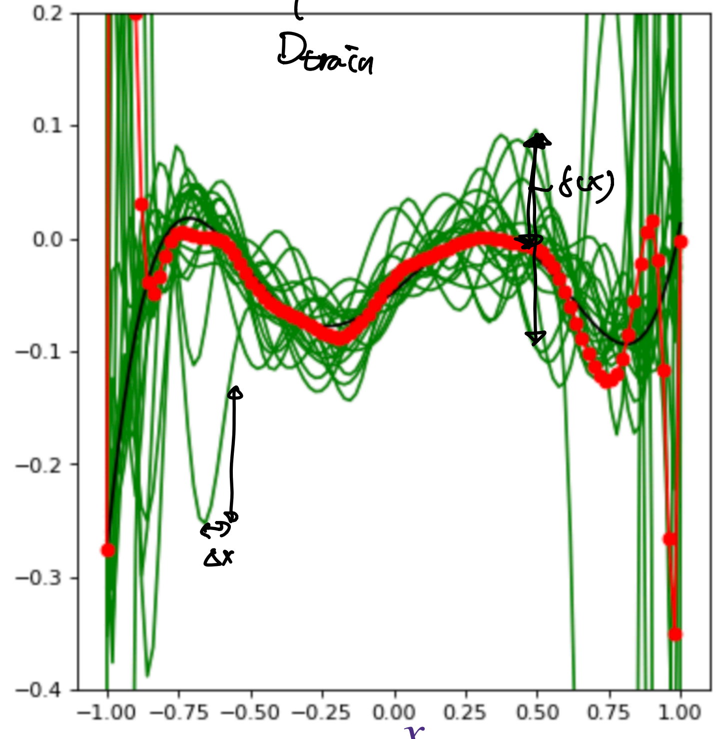

```
current train error = 0.0036791644380554187
current test error  = 0.0037962529988410953
```

With degree-20 polynomials, we overfit

bias $\downarrow$

$\hat{f}_{\hat{w}_{\mathrm{LS}}}(x)$

$\mathbb{E}_{D_{train}}[\delta^2(x)]$ ——— variance $\uparrow$




```
0.0005421686349568773
0.14210029429557927
```

Sensitivity: how to detect overfitting

- For a linear model,
  $y \simeq b + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$
  if $|w_j|$ is large then the prediction is sensitive to small changes in x_j
- Large sensitivity leads to overfitting and poor generalization, and equivalently models that overfit tend to have large weights
- Note that b is a constant and hence there is no sensitivity for the offset b

- In **Ridge Regression,** we use a regularizer $\underbrace{\|w\|_2^2}_{w_1^2+w_2^2+\cdots+w_d^2}$ to measure and control the sensitivity of the predictor
- And optimize for small loss and small sensitivity, by adding a **regularizer** in the objective (assume no offset for now)

$$\widehat{w}_{ridge} = \underbrace{\arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2}_{\text{Least-squares objective} = \text{fit}} + \lambda \|w\|_2^2$$

Regularization coefficient

$\lambda \geq 0$

Ridge Regression

$x_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $f(w) = w_1^2$, $w = (w_1, w_2)$

$X = \underset{n}{\begin{bmatrix} \overset{d}{-\, x_i^T \,-} \\ \vdots \end{bmatrix}}$

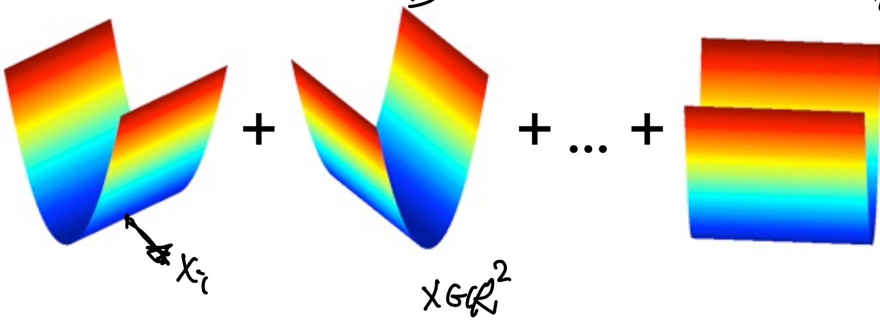
$x_i \sim N(0, I)$

$\mathbb{E}(x_i x_i^T) = I \xleftarrow{CLT} \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

$\boxed{n > d}$

- (Original) Least squares objective:

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

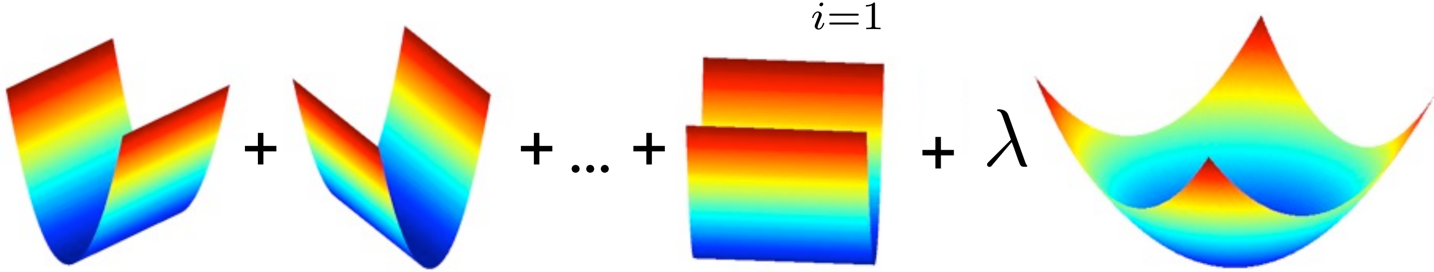


$(y_i - x_i^T w)^2 = f(w)$

$\boxed{w^T \underbrace{x_i x_i^T}_{\text{rank-1}} w} - 2 y_i x_i^T w = f(w)$

- Ridge Regression objective:

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



Minimizing the Ridge Regression Objective

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

$$\mathcal{L}(w) = ||Y - X \cdot w||_2^2 + \lambda ||w||_2^2$$

sometimes drop

$$\nabla_w \mathcal{L}(w) = -2X^T(Y - Xw) + 2\lambda w \Big|_{\hat{w}_{ridge}} = 0$$

$$X^T Y = (X^T X + \lambda \cdot \mathbb{I}) \cdot \hat{w}$$

$$\hat{w} = (X^T X + \lambda \cdot \mathbb{I})^{-1} \cdot X^T Y$$

Shrinkage Properties

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

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

Suppose $X^T X = n \cdot \mathbb{I}$

$\hat{w}_{LS}: \frac{1}{n} X^T Y$ $[\,] \in \mathbb{R}^d$, $\hat{w}_{ridge}: \frac{1}{n+\lambda} X^T Y$

$= \frac{n}{n+\lambda} \cdot \hat{w}_{LS}$

$\underbrace{\frac{n}{n+\lambda}}_{\text{Shrinking}}$

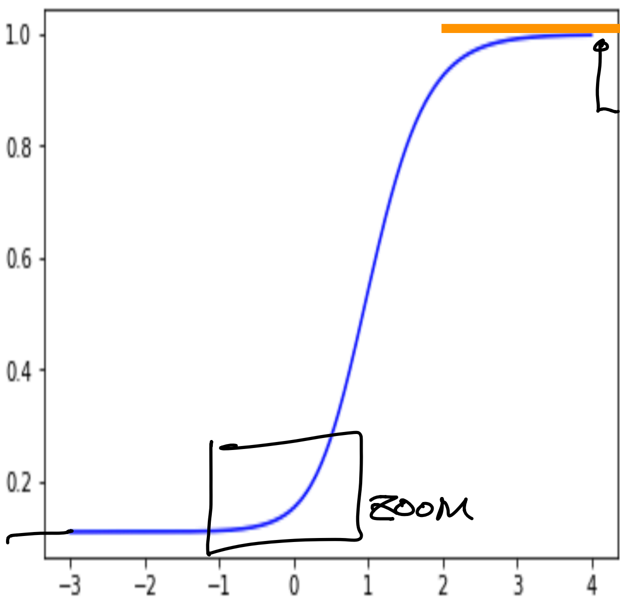
$X^T X \neq n \mathbb{I}$, in general.

- When $\lambda = 0$, this gives the least squares model
- This defines a family of models hyper-parametrized by λ
- Large λ means more regularization and simpler model
- Small λ means less regularization and more complex model

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

training MSE $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T \hat{w}^{(\lambda)}_{\text{ridge}})^2$

$\hat{w}_i$'s



$\frac{1}{n}\sum_{i=1}^{n} y_i^2$

$\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$

$=$

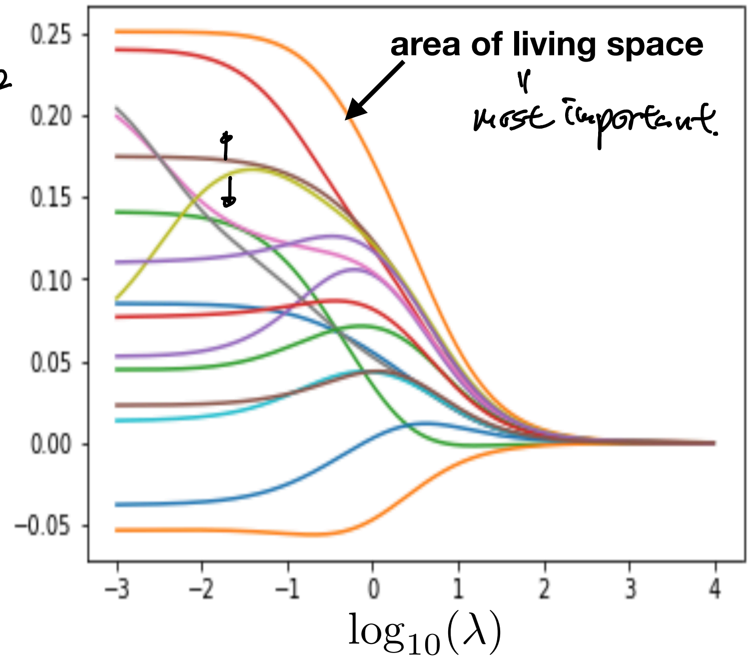
$\text{Var}(y_i)$

$\lambda \approx 0$

$\hat{w}_{\text{ridge}} = \hat{w}_{LS}$

$\lambda \approx \infty$

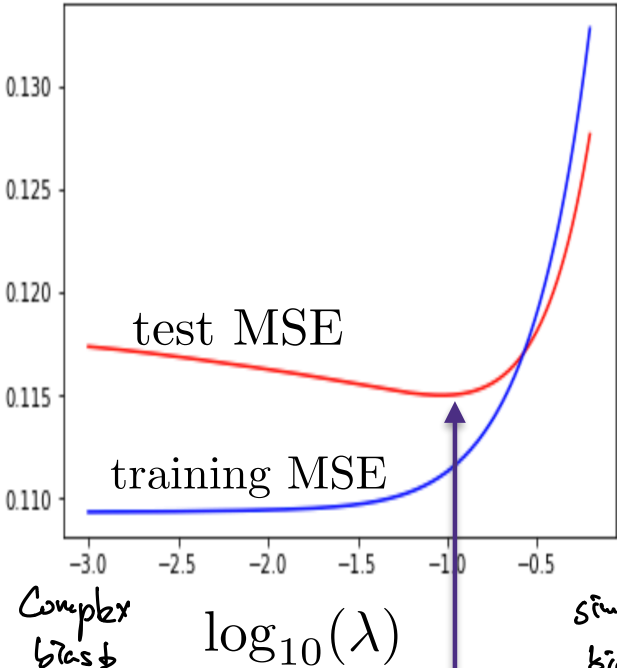
$\hat{w}_{\text{ridge}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

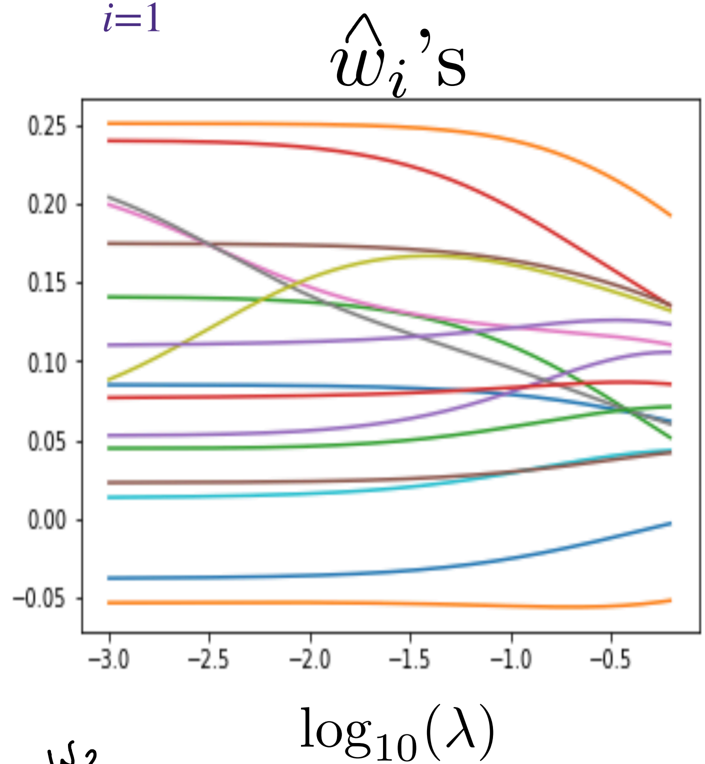


most important

- Left plot: leftmost training error is with no regularization: 0.1093
- Left plot: rightmost training error is variance of the training data: 0.9991
- Right plot: called **regularization path**

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

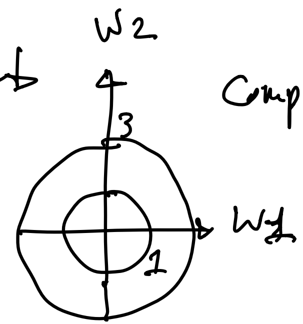




Complex
bias ↓
variance ↑

simple
bias ↑
variance ↓

Complexity ($\mathcal{F}$)
↳ class of predictors you search over.



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

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X, \;\; \mathbf{y} = \mathbf{X}w + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is
  $\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$ $\leftarrow$ best predictor $\mathbb{E}[Y|X]$

  $- \mathbb{E}[Y|X] + \mathbb{E}[Y|X]$

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}w + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

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

$\mathbb{E}[y|x] = x^T w$

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X,\ \ \mathbf{y} = \mathbf{X}w + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - \mathbb{E}[y\,|\,x])^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}[y\,|\,x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \underbrace{\mathbb{E}_{y|x}[(y - x^T w)^2 \,|\, x]}_{Y = x^T w + \varepsilon \quad \| \quad \mathbb{E}[\varepsilon^2] = \sigma^2} + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(x^T w - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X, \;\; \mathbf{y} = \mathbf{X}w + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - \mathbb{E}[y\,|\,x])^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}[y\,|\,x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - x^Tw)^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(x^Tw - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{(x^Tw - \mathbb{E}_{\mathcal{D}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x])^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}_{\tilde{\mathcal{D}}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]}_{\text{Variance}}$$

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X, \ \ \mathbf{y} = \mathbf{X}w + \epsilon, \ \ \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - \mathbb{E}[y\,|\,x])^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}[y\,|\,x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - x^T w)^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(x^T w - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{(x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x])^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}_{\tilde{\mathcal{D}}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]}_{\text{Variance}}$$

Suppose $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\overbrace{\mathbf{X}w + \epsilon}^{Y})$

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

$= \frac{1}{n+\lambda} \cdot I$

$$= \frac{n}{n+\lambda} w + \frac{1}{n+\lambda}\mathbf{X}^T \epsilon$$

Bias-Variance Properties

Suppose $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

$$\hat{w}_{\text{ridge}} = \frac{n}{n+\lambda} w + \frac{1}{n+\lambda}\mathbf{X}^T\epsilon$$

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X,\ \ \mathbf{y} = \mathbf{X}w + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - \mathbb{E}[y\,|\,x])^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}[y\,|\,x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \mathbb{E}_{y|x}[(y - x^T w)^2 \,|\, x] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(x^T w - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

$$= \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x])^2 + \mathbb{E}_{\mathcal{D}_{\text{train}}}[(\mathbb{E}_{\tilde{\mathcal{D}}_{\text{train}}}[x^T\hat{w}_{\text{ridge}} \,|\, x] - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$$

(verify at home)

$$= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{\frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2}_{\text{Bias-squared}} + \underbrace{\frac{\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2}_{\text{Variance}}$$

① $\lambda \uparrow \infty$: variance $\downarrow 0$
Bias $\uparrow (w^T x)^2$

② $\lambda \downarrow 0$: Bias $\downarrow 0$
variance $\uparrow \frac{\sigma^2\|x\|_2^2}{n}$

Bias-Variance Properties

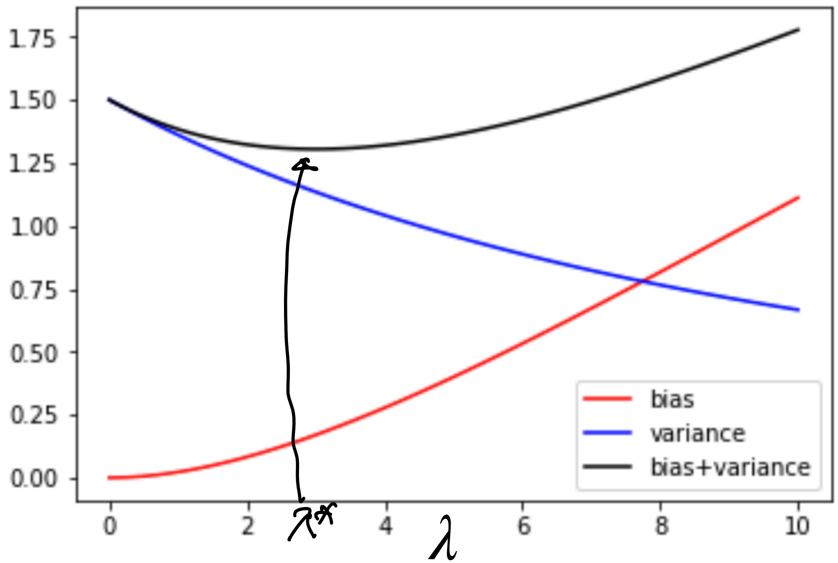
- Ridge regressor: $\widehat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$
- True error

$$\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T \hat{w}_{\text{ridge}})^2 \,|\, x] = \sigma^2 + \underbrace{\frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2}_{\text{Bias-squared}} + \underbrace{\frac{\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2}_{\text{Variance}}$$

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

as λ →0,

$\hat{w}_{\text{ridge}} \rightarrow \hat{w}_{\text{LS}}$



as λ →∞

$\hat{w}_{\text{ridge}} \rightarrow 0$

What you need to know…

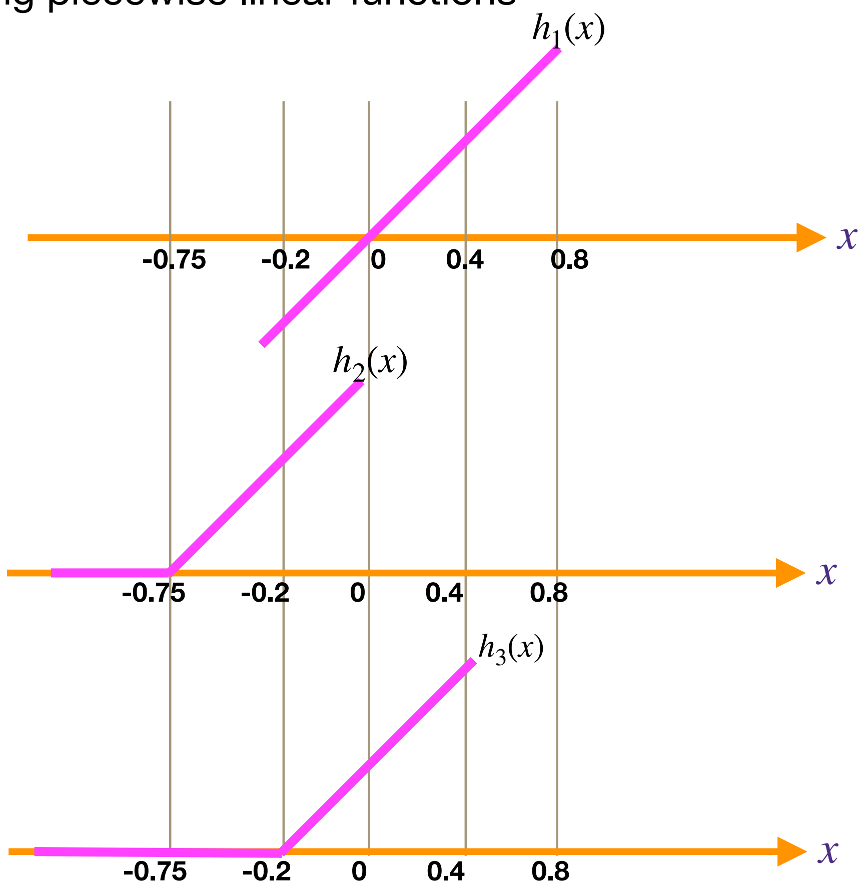
- Regularization
  - Penalizes complex models towards preferred, simpler models
- Ridge regression
  - L_2 penalized least-squares regression
  - Regularization parameter trades off model complexity with training error
  - Never regularize the offset!

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

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

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

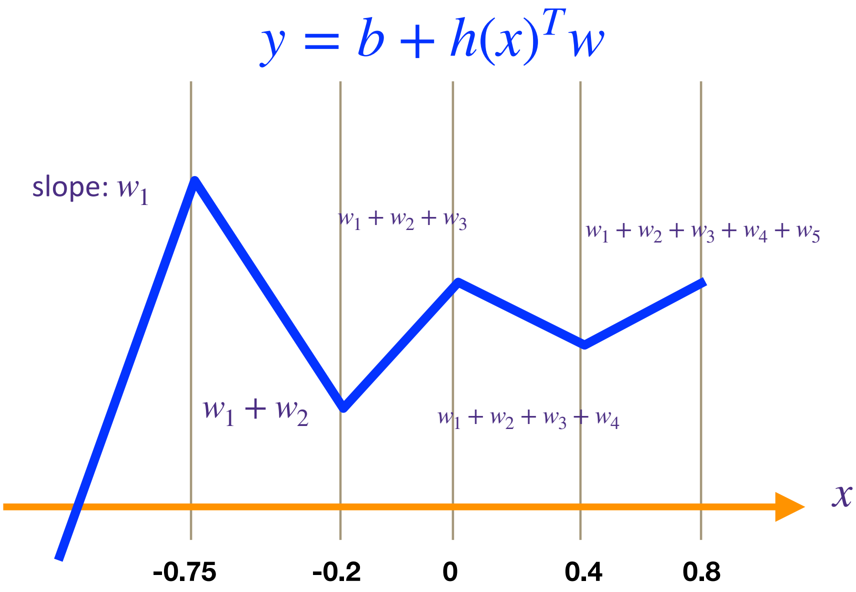


Example: piecewise linear fit

- 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

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

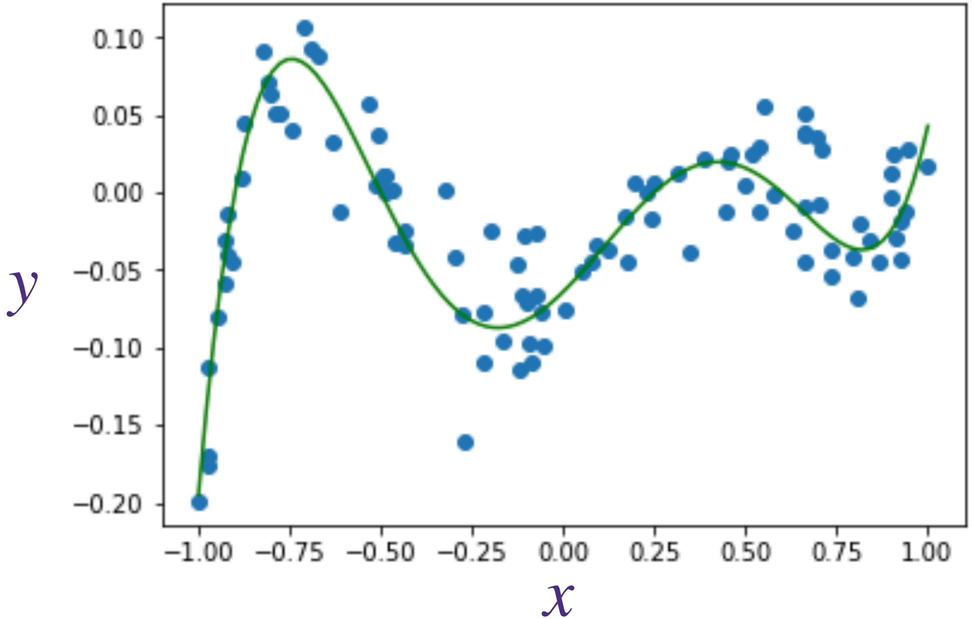
$$[a]^+ \triangleq \max\{a, 0\}$$



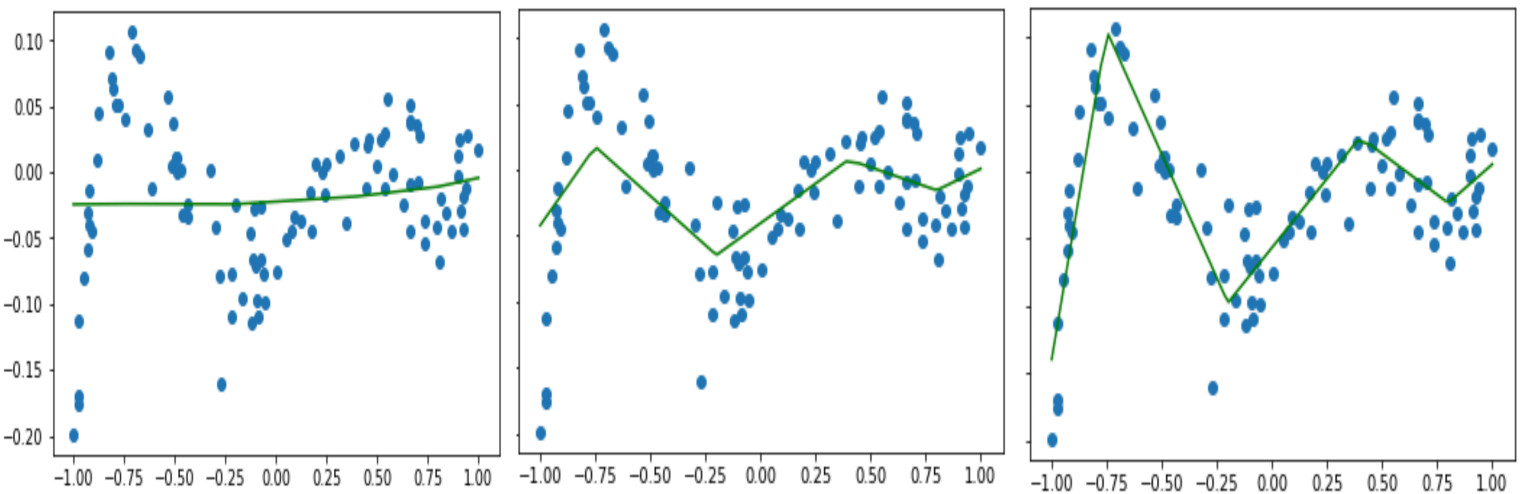
the weights capture the change in the slopes

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 << n=100

Piecewise linear with $w \in \mathbb{R}^{10}$ and n=11 samples