# Trading off bias and variance, Cross-validation

# Bias-variance tradeoff for least squares

If $Y_i = X_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\widehat{w}_{\text{MLE}} = w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = x^T w^*$$

$$\hat{f}_{\mathcal{D}}(x) = x^T w^* + x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

- Variance: $\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2\right] = \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$

$$= \sigma^2\,\mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$$

$$= \sigma^2\,x^T\mathbb{E}_{\mathcal{D}}[(\mathbf{X}^T\mathbf{X})^{-1}]x$$

- To analyze this, let's assume that $X_i \sim \mathcal{N}(0,\mathbf{I})$ and number of samples, $n$, is large enough such that $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$ with high probability and $\mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}] \simeq \frac{1}{n}\mathbf{I}$, then
  - Variance is $\frac{\sigma^2 x^T x}{n}$, and decreases with increasing sample size $n$

# Bias-Variance Properties of Ridge regression

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

# Bias-Variance Properties of Ridge regression

- 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

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

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

# Bias-Variance Properties of Ridge regression

- 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

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

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

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

# Bias-Variance Properties of Ridge regression

- 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 of Ridge regression

- 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

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

Irreduc. Error — Bias-squared — Variance

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

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

# 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)

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

Irreduc. Error — Bias-squared — Variance

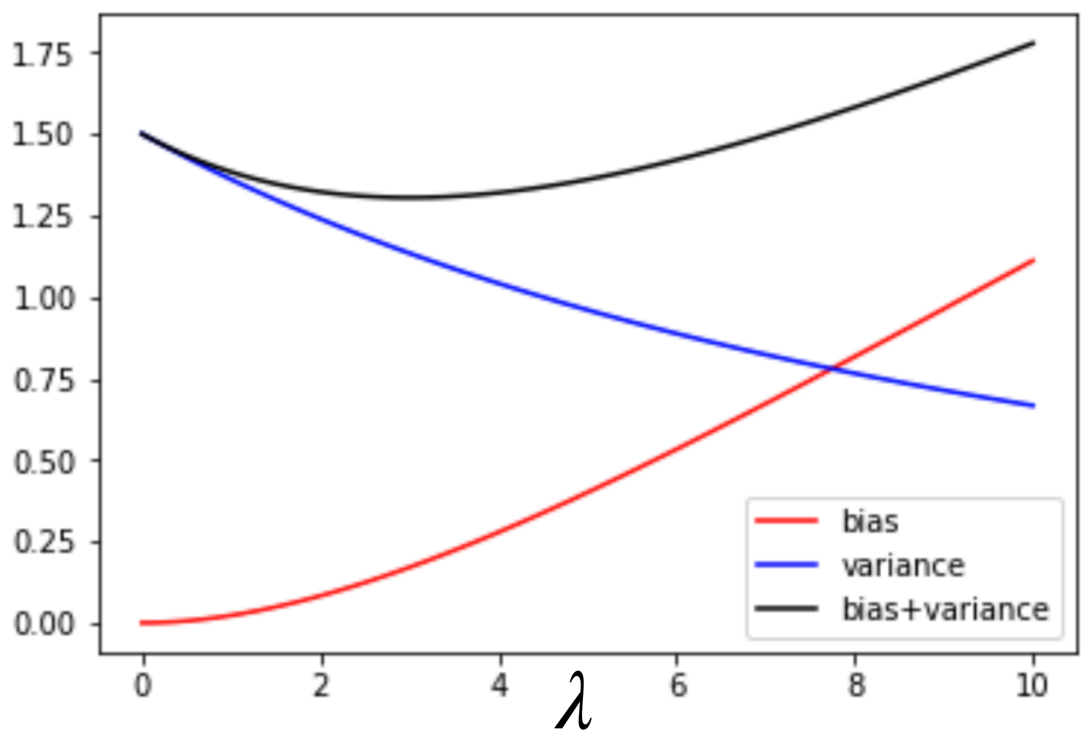
# Bias-Variance Properties of Ridge regression

- 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 $\lambda \to 0$, $\hat{w}_{\text{ridge}} \to \hat{w}_{\text{LS}}$



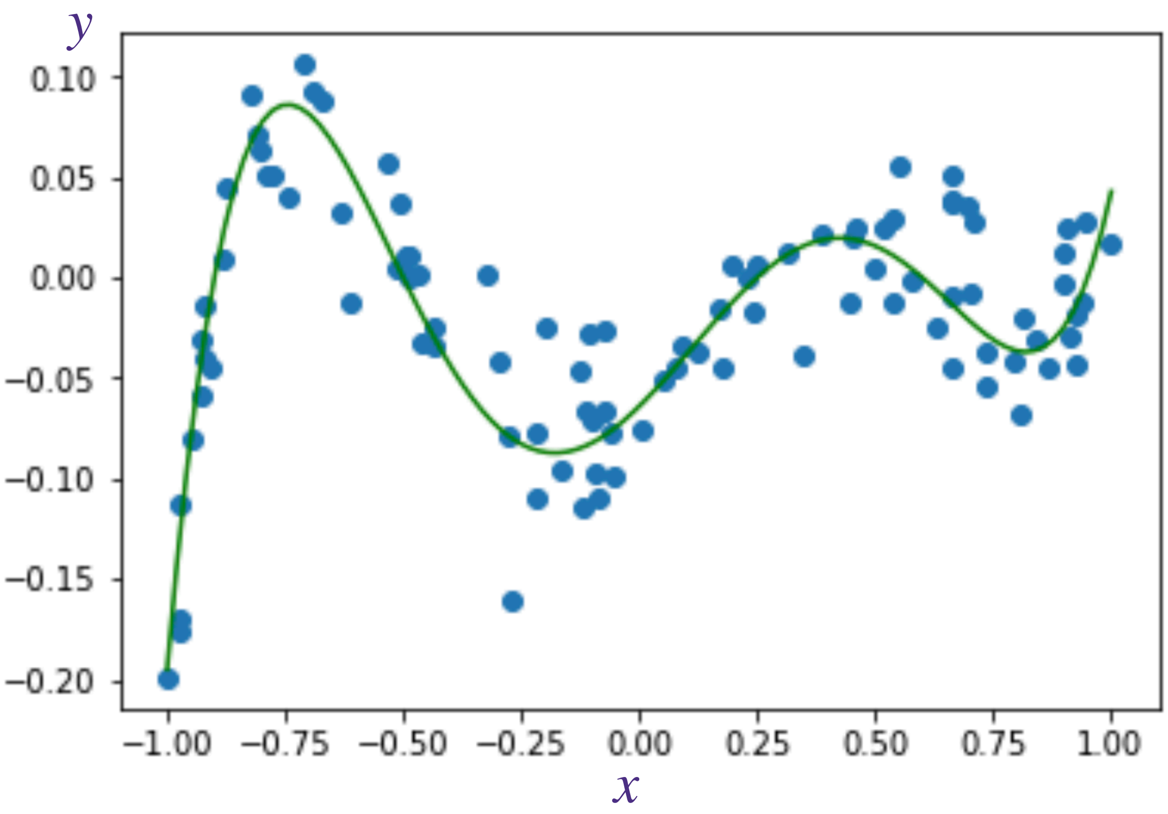

as $\lambda \to \infty$, $\hat{w}_{\text{ridge}} \to 0$

# 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

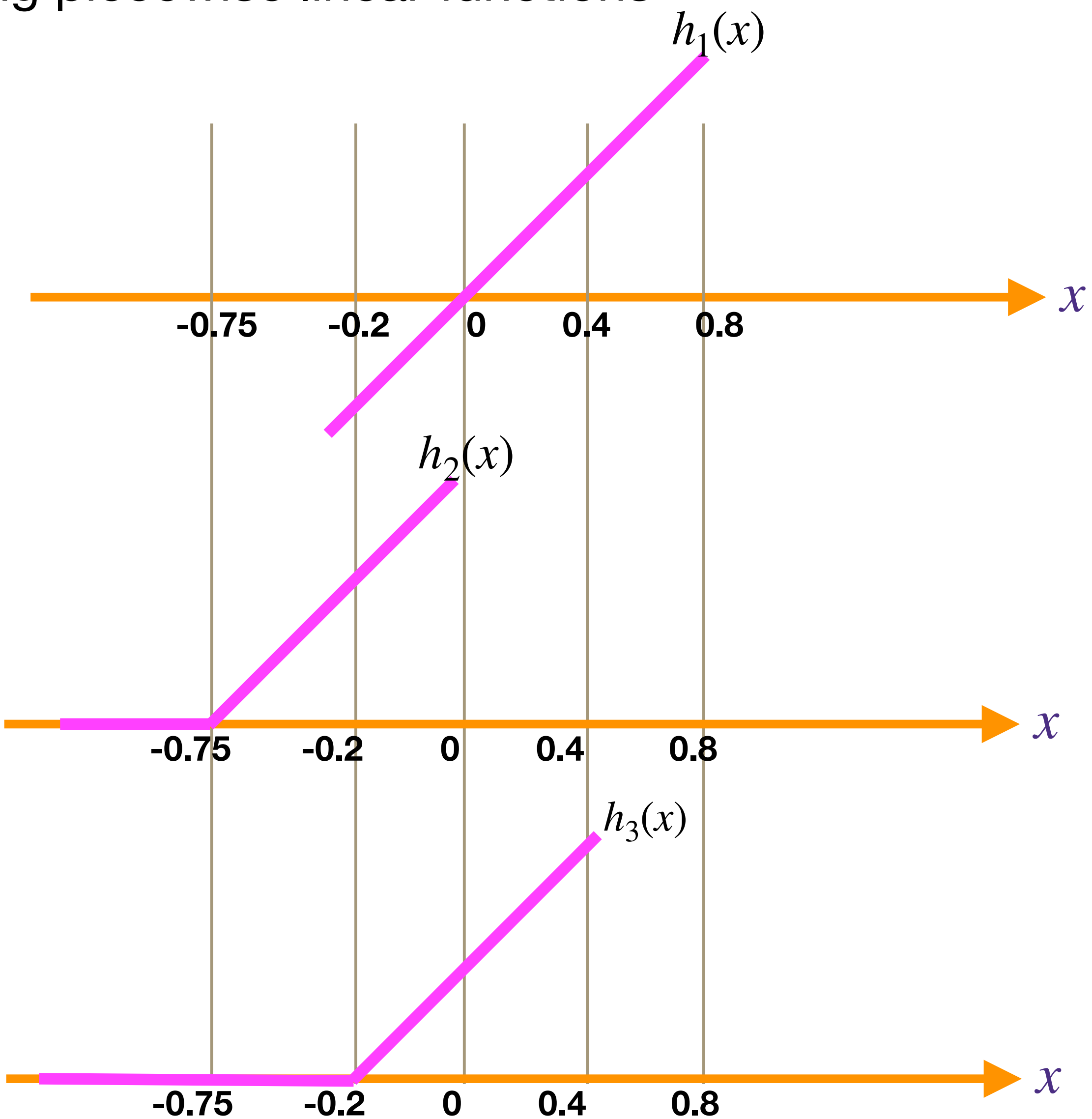


# Example: piecewise linear fit

- 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

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

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

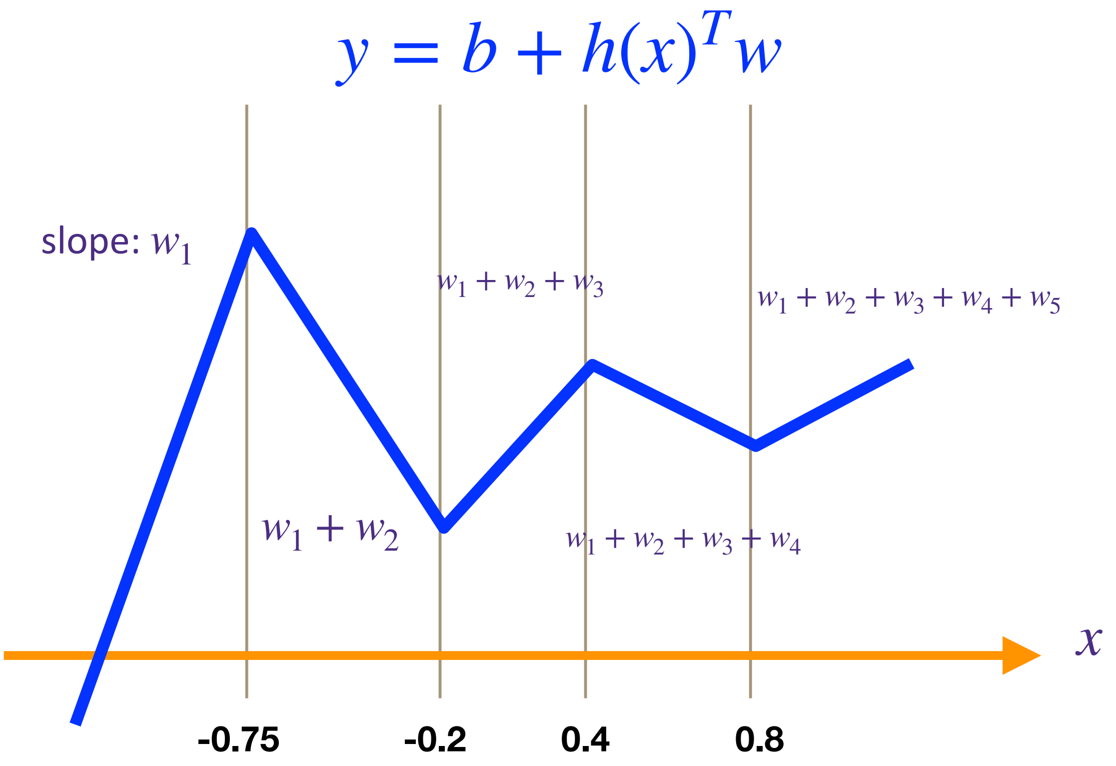


# Example: piecewise linear fit

- 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

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

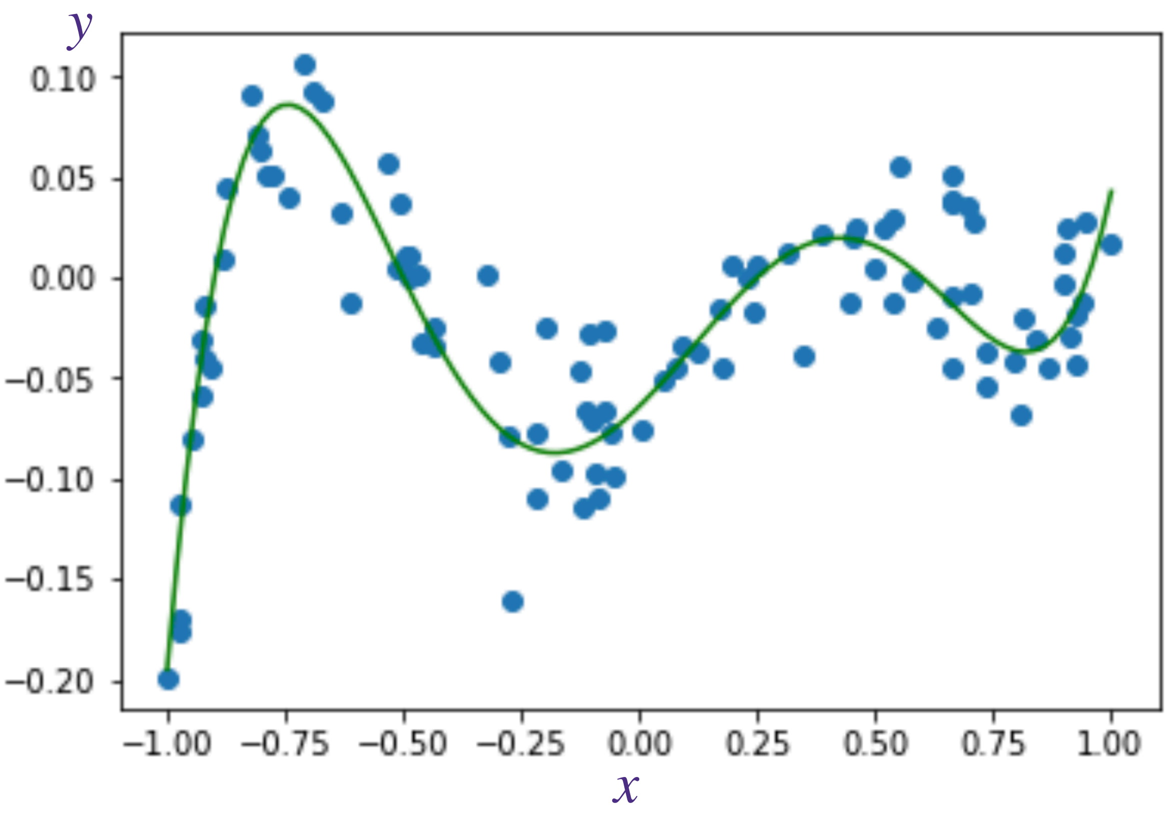
$$[a]^+ \triangleq \max\{a, 0\}$$



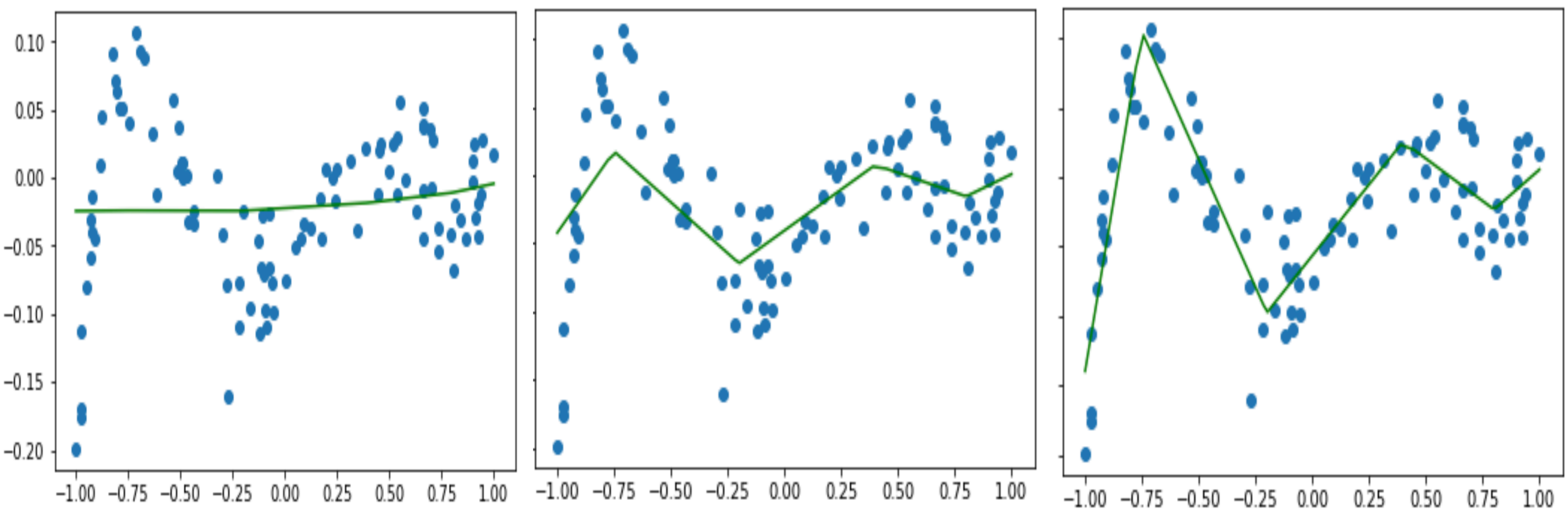

**the weights capture the change in the slopes**

# 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



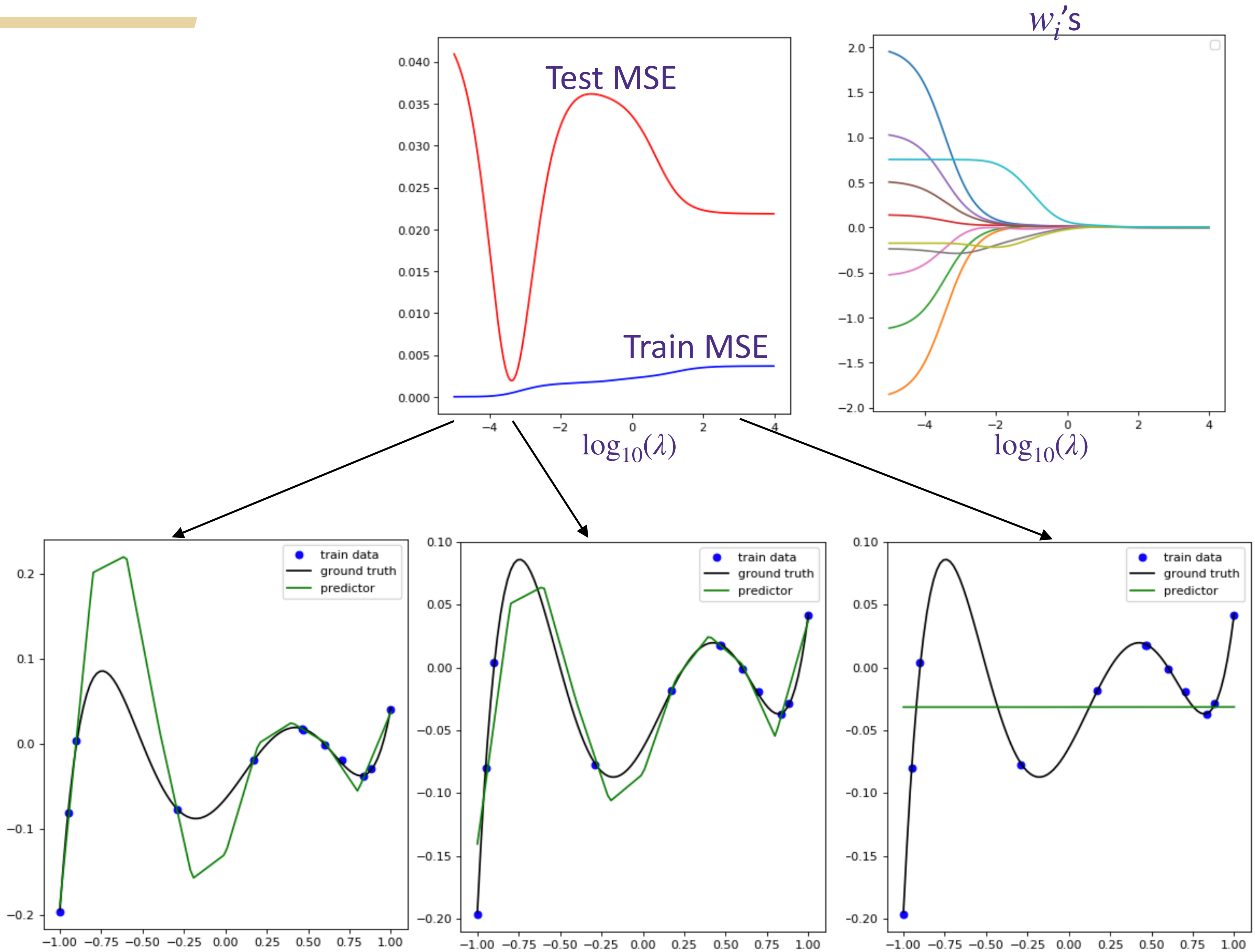
# Example: piecewise linear fit (ridge regression)




$\lambda = 1$ $\lambda = 0.005$ $\lambda = 0.000001$

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

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



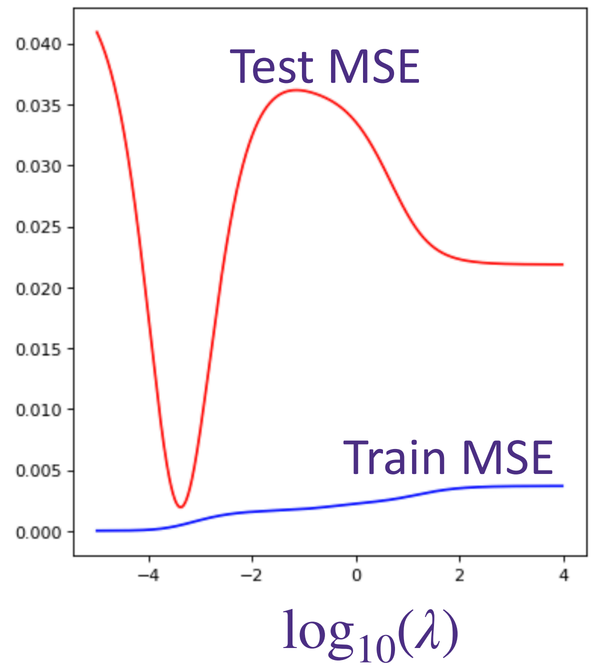
# Model selection using Cross-validation

# How… How… How????????

> Ridge regression:
How do we pick the regularization constant λ…

> Polynomial features:
How do we pick the number of basis functions…

> We could use the test data, but…



# How… How… How???????

- Ridge regression:
  How do we pick the regularization constant λ…
- Polynomial features:
  How do we pick the number of basis functions…

- We could use the test data, but…
  - Never ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever ever **train on the test data**

  - Use test data only for reporting the test error (once in the end)

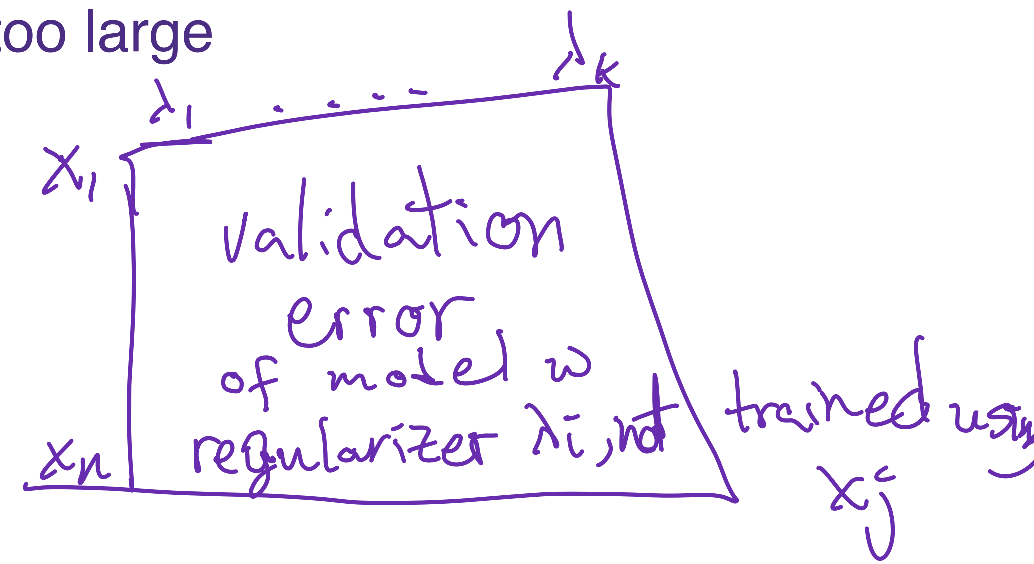
# (LOO) Leave-one-out cross validation

- Consider a validation set with 1 example:
  - $\mathscr{D}$ : training data
  - $\mathscr{D}\backslash j$ : training data with $j$-th data point $(x_j, y_j)$ moved to validation set
- Learn model $f_{\mathscr{D}\backslash j}$ with $\mathscr{D}\backslash j$ dataset
- The squared error on predicting $y_j$: $(y_j - f_{\mathscr{D}\backslash j}(x_j))^2$

  is an unbiased estimate of the **true error**

$$\text{error}_{\text{true}}(f_{\mathscr{D}\backslash j}) = \mathbb{E}_{(x,y)\sim P_{x,y}}[(y - f_{\mathscr{D}\backslash j}(x))^2]$$

  but, variance of $(y_j - f_{\mathscr{D}\backslash j}(x_j))^2$ is too large



# (LOO) Leave-one-out cross validation

> Consider a validation set with 1 example:
  - $\mathscr{D}$ : training data
  - $\mathscr{D}\backslash j$ : training data with $j$-th data point $(x_j, y_j)$ moved to validation set

> Learn model $f_{\mathscr{D}\backslash j}$ with $\mathscr{D}\backslash j$ dataset

> The squared error on predicting $y_j$: $(y_j - f_{\mathscr{D}\backslash j}(x_j))^2$
is an unbiased estimate of the **true error**

$$\text{error}_{\text{true}}(f_{\mathscr{D}\backslash j}) = \mathbb{E}_{(x,y)\sim P_{x,y}}[(y - f_{\mathscr{D}\backslash j}(x))^2]$$

but variance of $(y_j - f_{\mathscr{D}\backslash j}(x_j))^2$ is too large, so instead

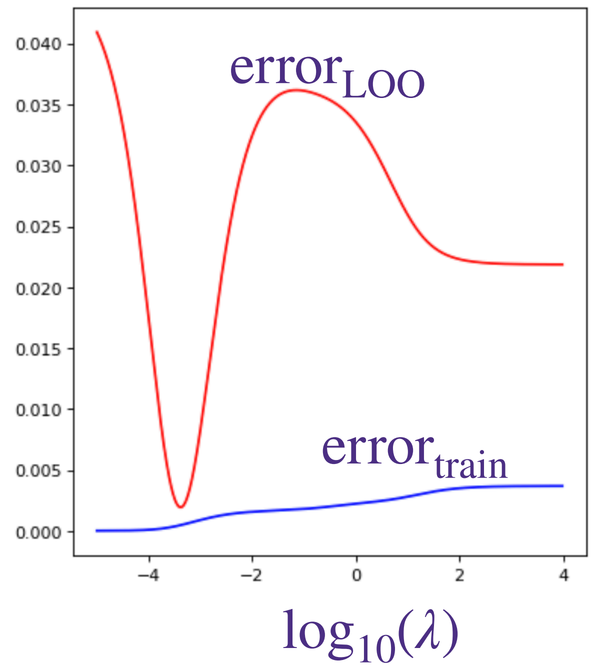
> **LOO cross validation**: Average over all data points $j$:
  - Train $n$ times:
    for each data point you leave out, learn a new classifier $f_{\mathscr{D}\backslash j}$
  - **Estimate the true error** as:

$$\text{error}_{LOO} = \frac{1}{n}\sum_{j=1}^{n}(y_j - f_{\mathcal{D}\backslash j}(x_j))^2$$

# LOO cross validation is (almost) unbiased estimate!

- When computing LOOCV error, we only use $n-1$data points to train
  - So it's not estimate of true error of learning with $n$ data points
  - Usually pessimistic – learning with less data typically gives worse answer. (Leads to an over estimation of the error)

- LOO is almost unbiased! Use LOO error for model selection!!!
  - **E.g., picking λ**



# Computational cost of LOO

- Suppose you have 100,000 data points
- say, you implemented a fast version of your learning algorithm
  - Learns in only 1 second
- Computing LOO will take about 1 day!!