Machine Learning

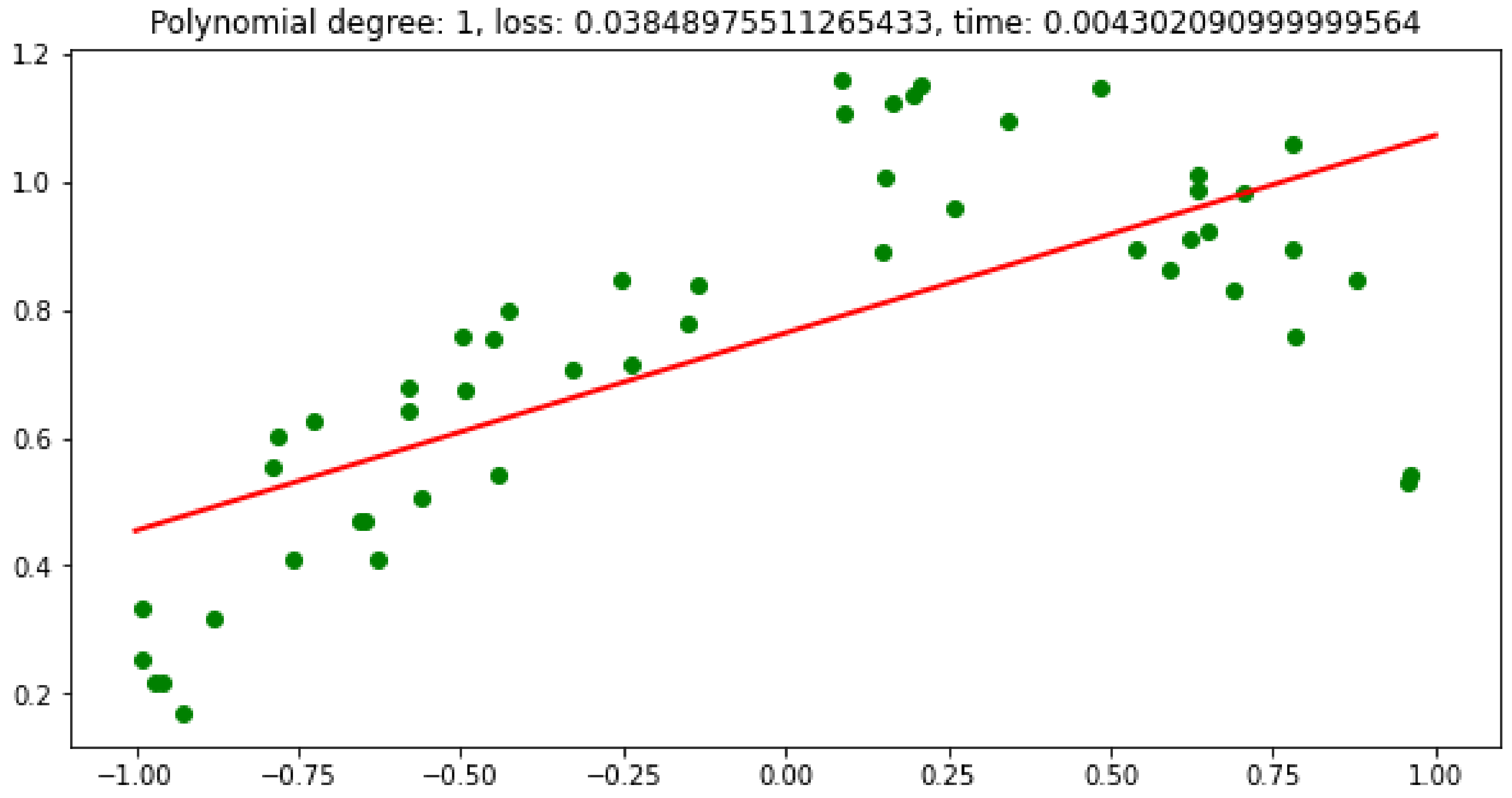
CMPT 726

Mo Chen
SFU School of Computing Science

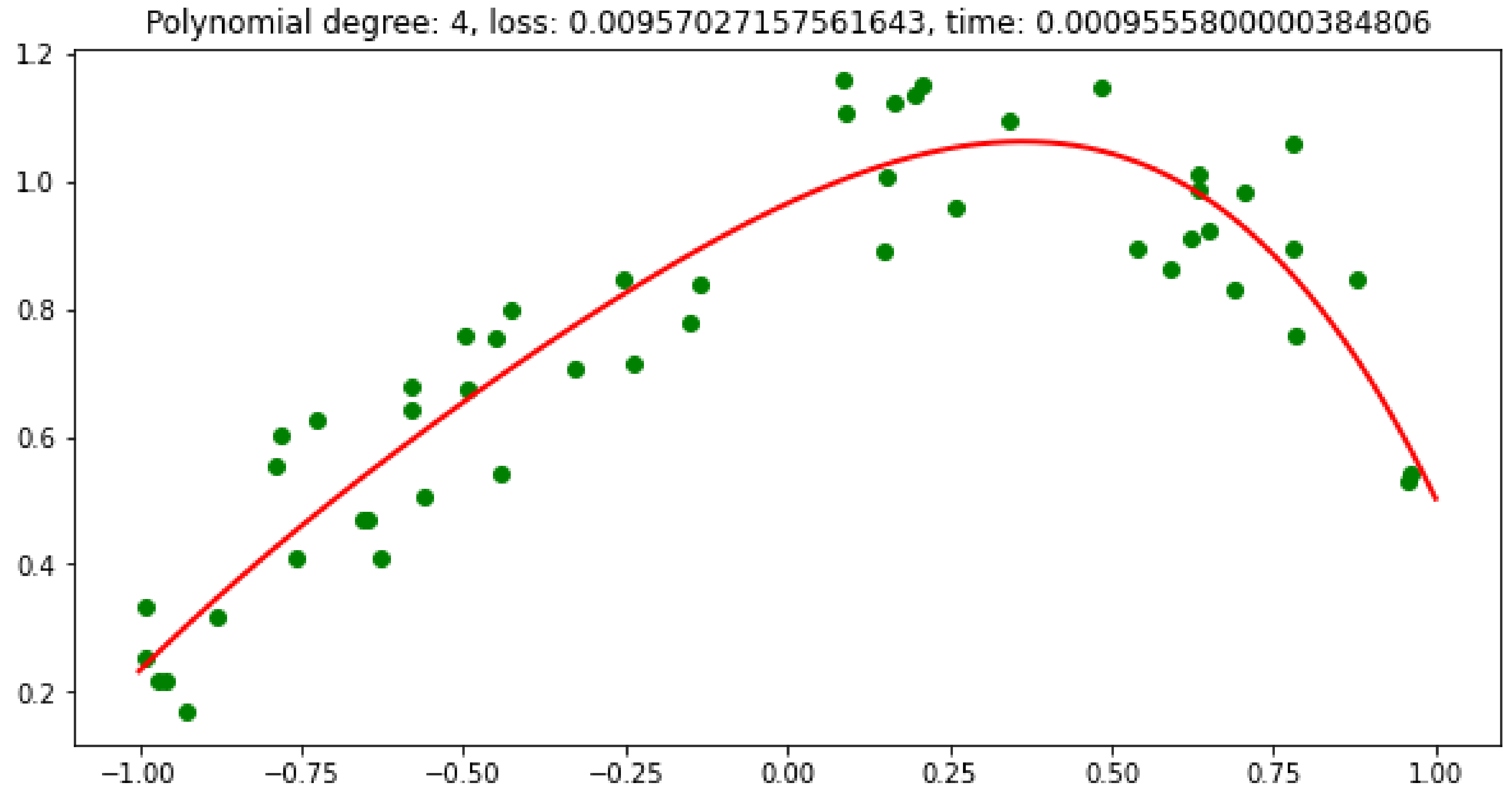
2022-09-29

Linear Regression (cont’d)

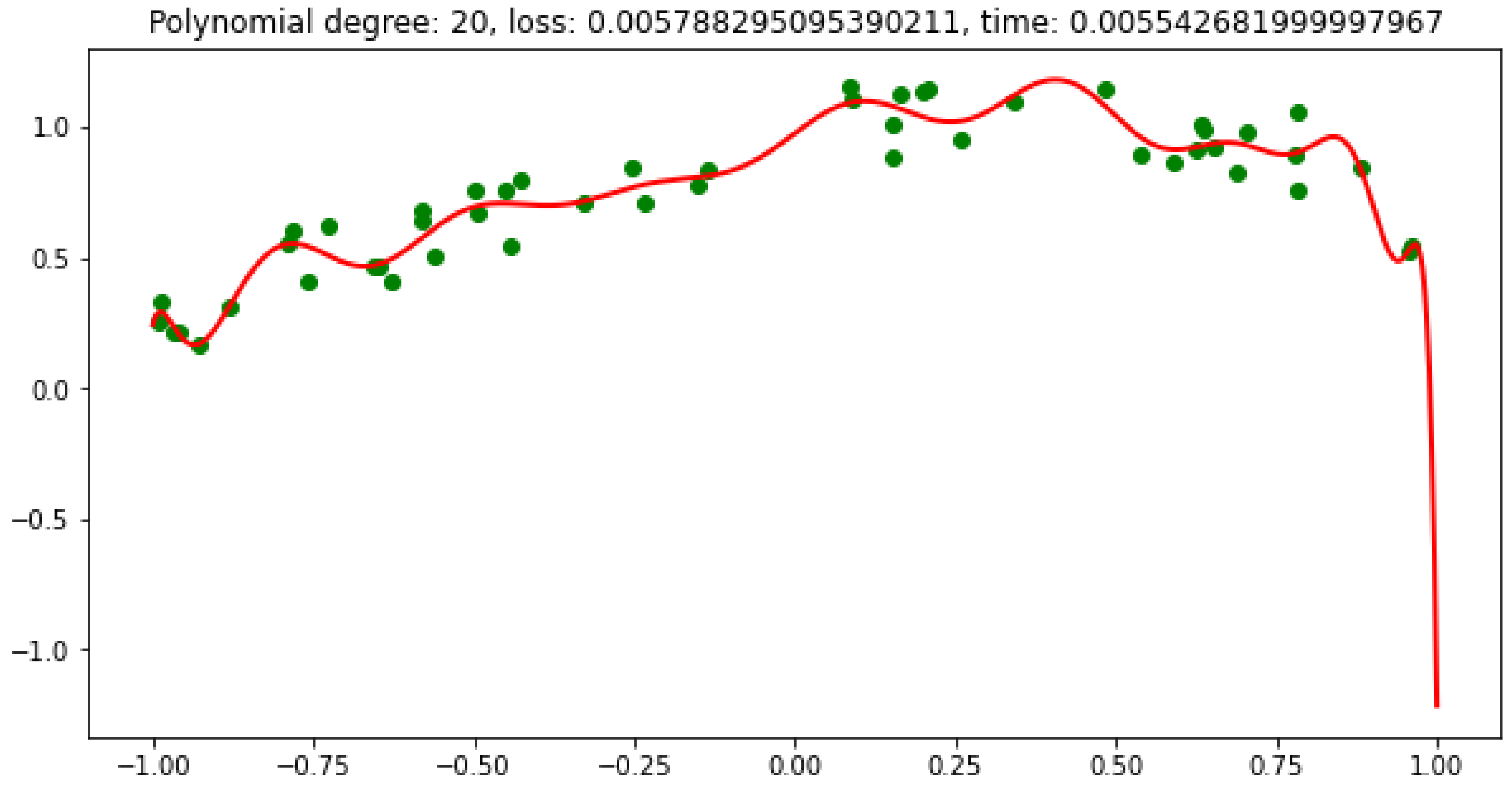
Overfitting vs. Underfitting


Polynomial degree: 1, loss: 0.03848975511265433, time: 0.004302090999999564

Underfitting


Polynomial degree: 4, loss: 0.00957027157561643, time: 0.0009555800000384806

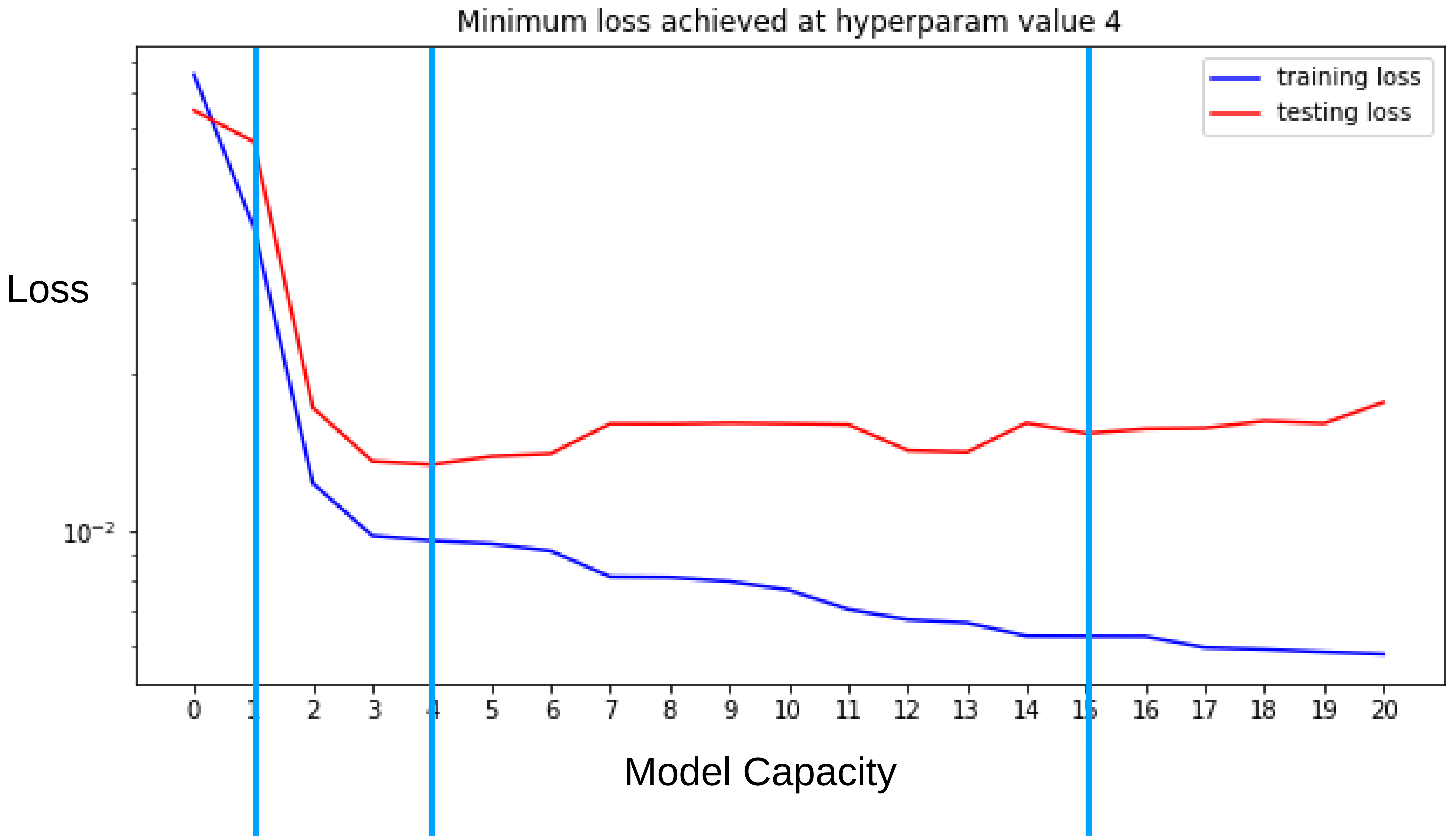
Just right


Polynomial degree: 20, loss: 0.005788295095390211, time: 0.005542681999997967

Overfitting

Overfitting vs. Underfitting

Just right



Loss

Model Capacity

Underfitting

Overfitting

Hyperparameters and Model Selection

Hyperparameters determine the model that is used to fit the data, e.g.: the degree of the polynomial

Choosing the best hyperparameter setting is known as **model selection**.

Never, ever perform model selection based on the testing loss!

Instead, split the training set into two subsets, a smaller training set and a **validation set**.

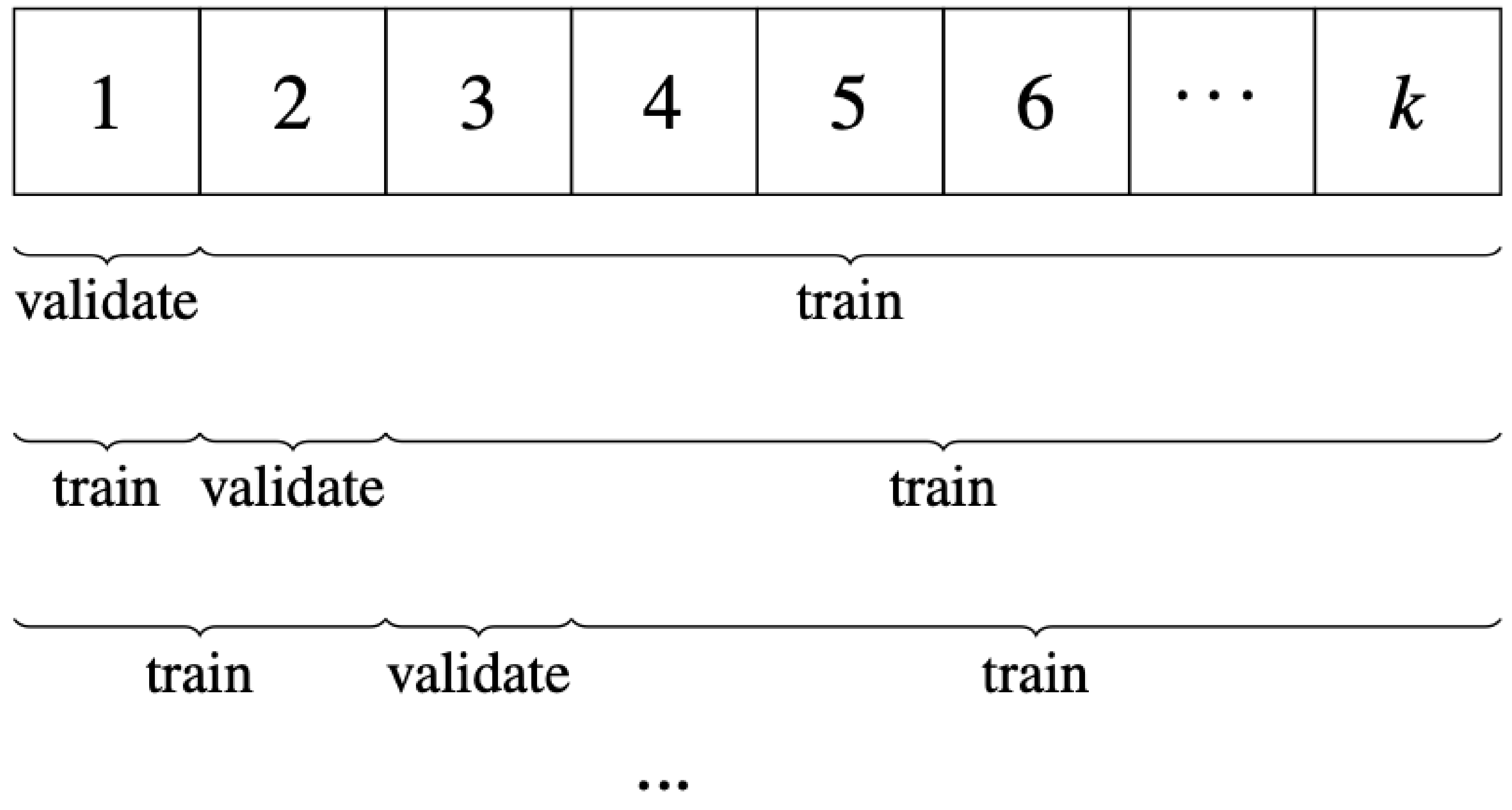
Validation set is not used for training and is only used for model selection.

K-Fold Cross Validation

In general, training a model on more data makes it perform better on **held-out data** (either validation or testing data).

Drawback of validation set: The training set is smaller, so the validation loss is a less accurate gauge of true performance on the testing set.

K-fold cross validation: Divide dataset into K equal-sized subsets, and train each model K times. Each time treat one of the subsets as the validation set and the others as the training set. At the end, average the K validation losses and use the average to perform model selection.



Useful when little data is available, but comes at the expense of greater computational cost.

Ridge Regression

Recall: When OLS overfits, $\vec{w}^*$ contains elements with large magnitude.

Idea: Change the loss function to penalize weights with large magnitude.

$$\text{OLS}: \quad L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 = \|\vec{y} - X\vec{w}\|_2^2$$

where $X = \begin{pmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_N^\top \end{pmatrix}, \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$

Ridge regression: $L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda \|\vec{w}\|_2^2 = \|\vec{y} - X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2$,

where $\lambda > 0$ is a hyperparameter.

Ridge Regression

$$L(\vec{w}) = \|\vec{y} - X\vec{w}\|_2^2 + \lambda\|\vec{w}\|_2^2$$
$$= (\vec{y} - X\vec{w})^\top(\vec{y} - X\vec{w}) + \lambda\vec{w}^\top\vec{w}$$
$$= \vec{y}^\top y - (X\vec{w})^\top\vec{y} - \vec{y}^\top(X\vec{w}) + (X\vec{w})^\top(X\vec{w}) + \lambda\vec{w}^\top\vec{w}$$
$$= \vec{y}^\top y - (2\vec{y}^\top X)\vec{w} + \vec{w}^\top(X^\top X)\vec{w} + \lambda\vec{w}^\top\vec{w}$$

$$\frac{\partial L}{\partial\vec{w}} = \frac{\partial(\vec{y}^\top\vec{y})}{\partial\vec{w}} - \frac{\partial\left((2X^\top\vec{y})^\top\vec{w}\right)}{\partial\vec{w}} + \frac{\partial(\vec{w}^\top(X^\top X)\vec{w})}{\partial\vec{w}} + \frac{\partial(\lambda\vec{w}^\top\vec{w})}{\partial\vec{w}} = 0$$
$$0 - 2X^\top\vec{y} + (X^\top X + (X^\top X)^\top)\vec{w} + \lambda(I + I^\top)\vec{w} = 0$$
$$-2X^\top\vec{y} + 2(X^\top X)\vec{w} + 2\lambda I\vec{w} = 0$$
$$-2X^\top\vec{y} + 2(X^\top X + \lambda I)\vec{w} = 0$$
$$2(X^\top X + \lambda I)\vec{w} = 2X^\top\vec{y}$$
$$(X^\top X + \lambda I)\vec{w} = X^\top\vec{y}$$
$$\vec{w} = (X^\top X + \lambda I)^{-1}X^\top\vec{y}$$

Ridge Regression

Recall: $\frac{\partial L}{\partial \vec{w}} = -2X^\top \vec{y} + 2(X^\top X + \lambda I)\vec{w}$

$$\frac{\partial^2 L}{\partial \vec{w} \partial \vec{w}^\top} = \frac{\partial}{\partial \vec{w}}\left(\frac{\partial L}{\partial \vec{w}}\right)$$

$$= \frac{\partial}{\partial \vec{w}}(-2X^\top \vec{y} + 2(X^\top X + \lambda I)\vec{w})$$
$$= \frac{\partial}{\partial \vec{w}}(-2X^\top \vec{y}) + \frac{\partial}{\partial \vec{w}}(2(X^\top X + \lambda I)\vec{w})$$
$$= 0 + 2(X^\top X + \lambda I)^\top$$
$$= 2(X^\top X + \lambda I)$$

Claim: $X^\top X + \lambda I \succ 0$

Proof: $\vec{w}^\top (X^\top X + \lambda I)\vec{w} = \vec{w}^\top (X^\top X)\vec{w} + \vec{w}^\top (\lambda I)\vec{w} = (X\vec{w})^\top (X\vec{w}) + \lambda \vec{w}^\top \vec{w} = \|X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2$

For any $\vec{w} \neq \vec{0}$, $\|\vec{w}\|_2^2 > 0$

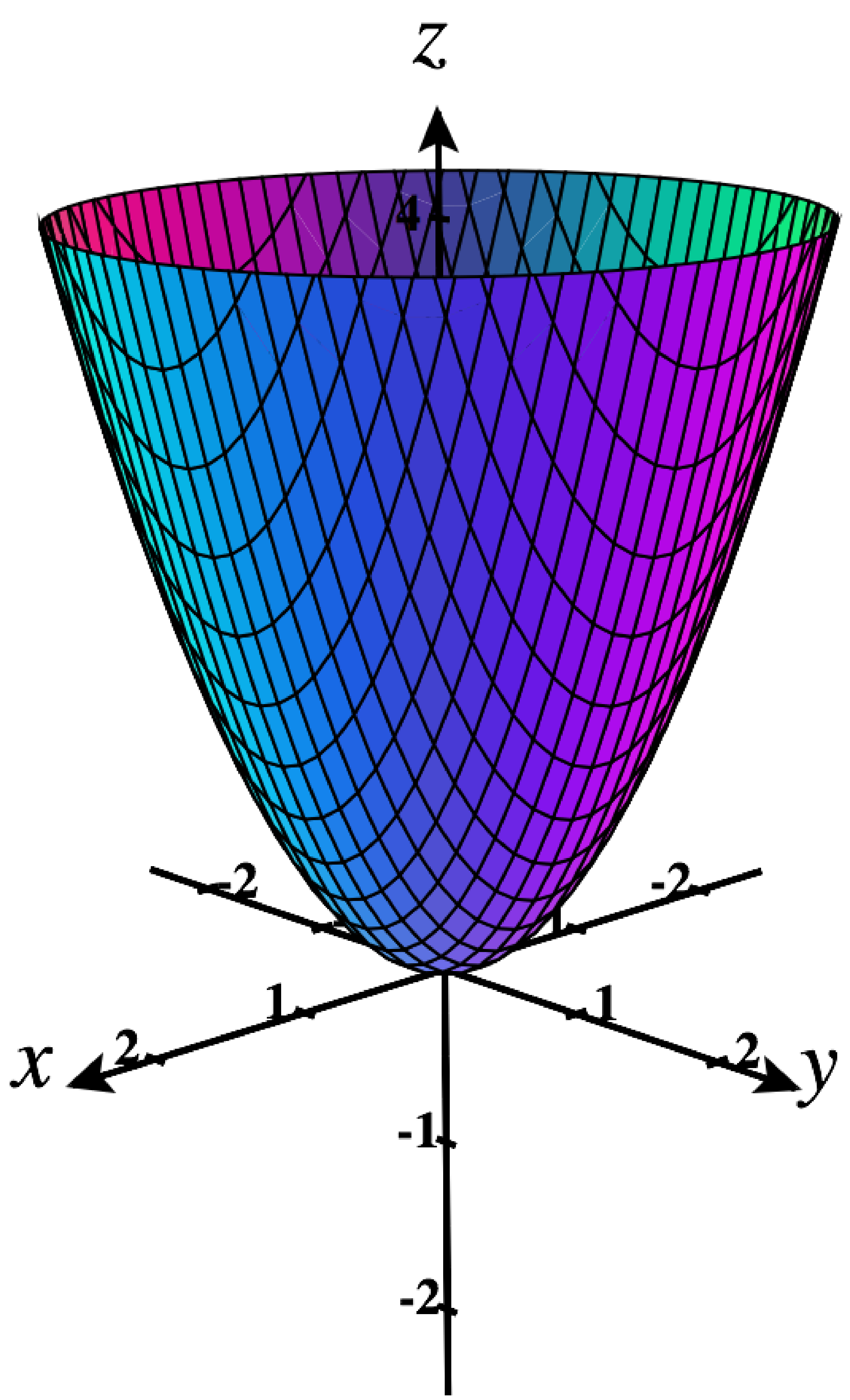
Since $\|X\vec{w}\|_2^2 \geq 0$ and $\lambda > 0$, $\|X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2 > 0 \quad \forall \vec{w} \neq \vec{0}$

So the loss function is strictly convex.

Ridge Regression

For a strictly convex function, there is a unique critical point, which is a local minimum, which is a global minimum.

So, the critical point $\vec{w}^* = (X^\top X + \lambda I)^{-1} X^\top \vec{y}$ is the only optimal parameter vector, regardless of whether X is full-rank or not.



Ridge Regression: Summary

Model: $\hat{y} = \vec{w}^\top \vec{x}$

Parameters: $\vec{w}$

Loss function: $L(\vec{w}) = \sum_{i=1}^{N}(y_i - \vec{y}_i)^2 + \lambda\|\vec{w}\|_2^2 = \sum_{i=1}^{N}(y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda\|\vec{w}\|_2^2$

where $X = \begin{pmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_N^\top \end{pmatrix}, \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$

Optimal parameters: $\vec{w}^* \coloneqq \arg\min_{\vec{w}} L(\vec{w}) = (X^\top X + \lambda I)^{-1} X^\top \vec{y}$

OLS vs. Ridge Regression

Model: $\vec{y} = \vec{w}^\top \vec{x}$; Parameters: $\vec{w}$

OLS:

Loss function:

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2$$
$$= \|\vec{y} - X\vec{w}\|_2^2$$

Optimal Parameters:

$$\vec{w}^* := \arg\min_{\vec{w}} L(\vec{w}) = (X^\top X)^{-1} X^\top \vec{y}$$

Ridge Regression:

Loss function:

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda \|\vec{w}\|_2^2$$
$$= \|\vec{y} - X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2$$

Optimal Parameters:

$$\vec{w}^* := \arg\min_{\vec{w}} L(\vec{w}) = (X^\top X + \lambda I)^{-1} X^\top \vec{y}$$