

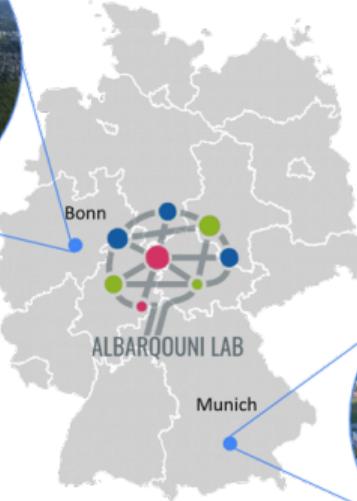
MACHINE LEARNING

Linear Models: Linear Regression

Last Update: 2nd November 2022

Prof. Dr. Shadi Albarqouni

Director of Computational Imaging Research Lab. (Albarqouni Lab.)
University Hospital Bonn | University of Bonn | Helmholtz Munich



**HELMHOLTZ
MUNICH**

STRUCTURE

1. Linear Regression

1.1 Terminology

1.2 Least Norm Estimation

1.3 Least Squares Estimation

1.4 Algorithmic issues

1.5 Measuring goodness of fit

LINEAR REGRESSION



LINEAR REGRESSION

Predicting Manhattan Rent with Linear Regression

An Analysis of StreetEasy Rental Listings



Photo by Florian Wehde on [Unsplash](#)

LINEAR REGRESSION -- HYPOTHESIS REPRESENTATION

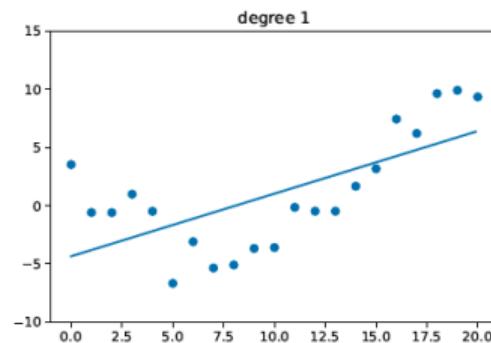
Definition

Linear regression is a widely used regression model $p(y|\mathbf{x}; \theta) = \mathcal{N}(y|\mathbf{w}^T \mathbf{x} + b, \sigma^2)$ for predicting a real-valued output $\mathbf{y} \in \mathbb{R}$, given a fixed-dimensional input vector $\mathbf{x} \in \mathbb{R}^D$ (also called independent variables, explanatory variables, or covariates) where $\theta = (\mathbf{w}, b, \sigma^2)$ are the parameters with \mathbf{w} as **weights** or **regression coefficients** and b or w_0 as the **offset** or **bias** term.

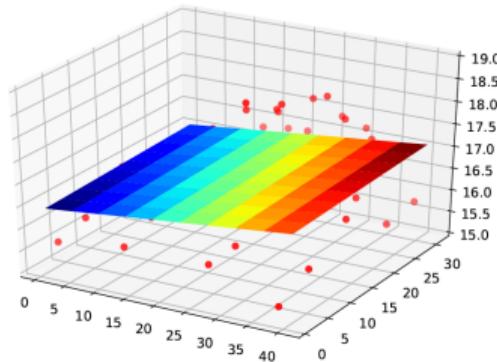
The key property of the model is that the expected value of the output is assumed to be a linear function of the input, $\mathbb{E}[y|x] = w^T x$, which makes the model easy to interpret, and easy to fit to data.

TERMINOLOGY

Simple linear regression: The input is one-dimensional (so $D = 1$), the model has the form $f(x; w) = ax + b$, where $b = w_0$ is the **intercept**, and $a = w_1$ is the **slope**.



Multiple linear regression: The input is multi-dimensional, $x \in \mathbb{R}^D$ where $D > 1$.

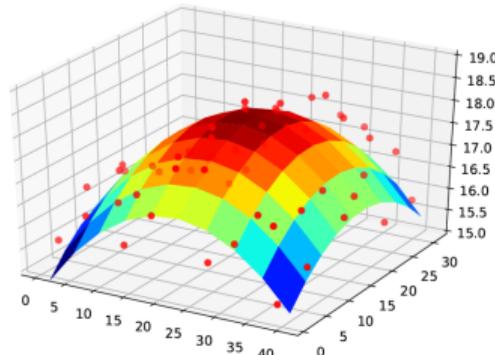
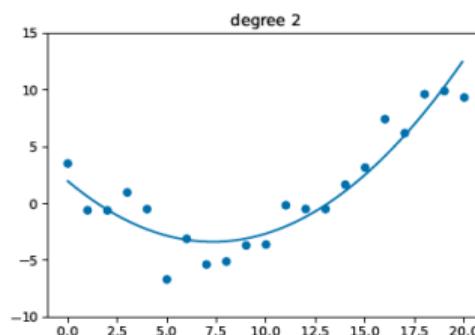


TERMINOLOGY -- CONT.

Multivariate linear regression: The output is multi-dimensional, $y \in \mathbb{R}^J$ where $J > 1$, and the likelihood can be written as

$$p(y|\mathbf{x}; \theta) = \prod_{j=1}^J \mathcal{N}(y_j | \mathbf{w}_j^T \mathbf{x}, \sigma_j^2)$$

Polynomial linear regression: A non-linear transformation $\phi(\cdot)$, e.g., a polynomial expansion of degree d is applied to the input vector. Consider a one-dimensional input (so $D = 1$), the $\phi(x) = [1, x, x^2, \dots, x^d]$ and the likelihood can be written as $p(y|\mathbf{x}; \theta) = \mathcal{N}(y | \mathbf{w}^T \phi(\mathbf{x}), \sigma^2)$



Polynomial Linear Regression in for 1D and 2D inputs

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Maximum Likelihood Estimation (MLE)

It can be obtained by minimizing the Negative Log Likelihood as an objective function

$$\theta_{MLE} = \arg \min_{\theta} NLL(\theta) \quad \text{where} \quad \theta = (\mathbf{w}, b, \sigma^2) \triangleq (\mathbf{w}, \sigma^2)^1$$

The **Negative Log Likelihood (NLL)** for the linear regression is given by

$$NLL(\theta) = -\log \prod_{n=1}^N \underbrace{\mathcal{N}(y_n | w^T x_n + b, \sigma^2)}_{p(y_n | x_n; \theta)} \triangleq \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - \hat{y}_n)^2 + \frac{N}{2} \log(2\pi\sigma^2)$$

where $\hat{y}_n = f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n$ is the prediction with bias $w_0 = b$ and $x_0 = 1$.

The **NLL** is equal (up to irrelevant constants) to the **residual sum of squares**,

$$\text{RSS}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y_n - \hat{y}_n)^2 = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

¹ b is included in w by simply adding a column with a value of 1 to the feature vector $x_{1:D+1} = [1, x_{1:D}]$

LINEAR REGRESSION AS SYSTEMS OF EQUATIONS

Linear regression problem as *systems of equations*

$$y_1 = w_0 + w_1 x_{11} + \cdots + w_D x_{1D}$$

$$y_2 = w_0 + w_1 x_{21} + \cdots + w_D x_{2D}$$

...

$$y_N = w_0 + w_1 x_{N1} + \cdots + w_D x_{ND}$$

The system of equations can be written in a matrix form as $y = Xw$ with $Y \in \mathbb{R}^N$ as targets, $X \in \mathbb{R}^{N \times D}$ as design input matrix, and $w \in \mathbb{R}^D$ as the weight parameters.

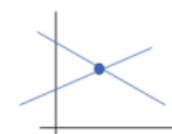
if $N < D$, the system is **underdetermined**, so there is **not a unique solution** → the minimal norm solution is demonstrated

if $N = D$ and w is full rank, there is a single unique solution

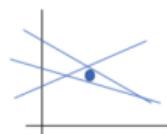
if $N > D$, the system is **overdetermined**, so there is **no unique solution** → the least square solution is demonstrated



$N < D$



$N = D$



$N > D$

Machine Learning

- Linear Regression
- Terminology
 - Linear Regression as Systems of Equations

Given the following systems of equations,

$$\begin{aligned} 2 &= 3w_1 + 2w_2 \\ -2 &= 2w_1 - 2w_2 \end{aligned}$$

it can be written in the matrix form $y = Xw$ as

$$y = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \end{pmatrix}$$

Since $N = D$, we can simply solve the systems using

$$w = X^{-1}y = \frac{1}{|detX|} \begin{pmatrix} -2 & -2 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

What happens if:

we have only the first data point (equation)?

we have an additional data point $0 = -w_1 + 3w_2$?

LINEAR REGRESSION AS SYSTEMS OF EQUATIONS

Linear regression problem as **systems of equations**:

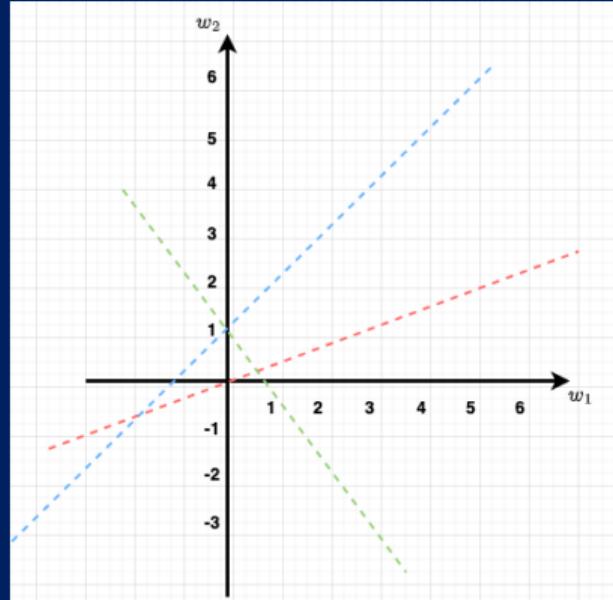
$$\begin{aligned} y_0 &= w_0 + w_1x_1 + \dots + w_Dx_D \\ y_1 &= w_0 + w_1x_1 + \dots + w_Dx_D \\ &\dots \\ y_N &= w_0 + w_1x_1 + \dots + w_Dx_D \end{aligned}$$

The system of equations can be written in a matrix form as $y = Xw$ with $y \in \mathbb{R}^N$ as targets, $X \in \mathbb{R}^{N \times D}$ as design input matrix, and $w \in \mathbb{R}^D$ as the weight parameters.

if $N < D$, the system is underdetermined, so there is **not a unique solution** → the minimal norm solution is demonstrated

if $N = D$ and w is full rank, there is a **single unique solution**

if $N > D$, the system is over-determined, so there is **no unique solution** → the least square solution is demonstrated



LEAST NORM ESTIMATION

When $N < D$ (**short and fat**), the system is **underdetermined**, so there is **not** a unique solution → Least norm estimation?

Least Norm Estimation

$$\hat{w} = \arg \min_w \|w\|_2^2 \quad \text{s.t.} \quad Xw = y$$

The minimal norm solution is obtained using the **right pseudo inverse**:

$$w_{pinv} = X^T \underbrace{(X X^T)^{-1}}_{\mathbb{R}^{N \times N}} y$$

Proof → Have a look at Sec. 7.7.2 and Sec. 7.5.3

Machine Learning

└ Linear Regression

└ Least Norm Estimation

└ Least Norm Estimation

What happens if:

we have only the first data point (equation)?

Given the following systems of equations,

$$2 = 3w_1 + 2w_2$$

it can be written in the matrix form $y = Xw$ as

$$y = 2, \quad X = \begin{pmatrix} 3 & 2 \end{pmatrix}$$

Since $N < D$, we can simply solve the systems using

$$w_{pinv} = X^T(XX^T)^{-1}y = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \left(\begin{pmatrix} 3 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right)^{-1} (2) = \begin{pmatrix} 0.46 \\ 0.31 \end{pmatrix}.$$

LEAST NORM ESTIMATION

When $N < D$ (**short and fat**), the system is **underdetermined**, so there is **not a unique solution** → Least norm estimation?

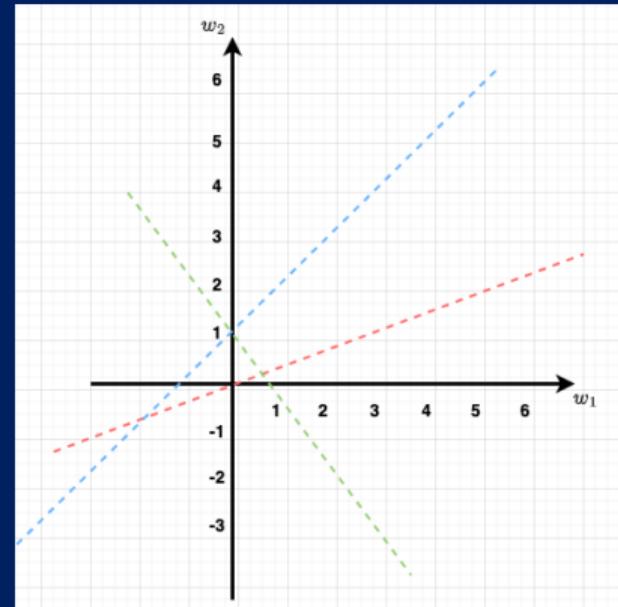
Least Norm Estimation

$$\hat{w} = \arg \min_w \|w\|_F^2 \quad \text{s.t.} \quad Xw = y$$

The minimal norm solution is obtained using the **right pseudo inverse**:

$$w_{pinv} = X^T \frac{(XX^T)^{-1}y}{\|y\|_F^2}$$

Proof → Have a look at Sec. 7.7.2 and Sec. 7.5.3



LEAST SQUARES ESTIMATION

When $N > D$ (tall and skinny), the system is **overdetermined**, so there is no unique solution → Least square estimation?

Least Squares Estimation (LSE)

To find the solution that gets as close as possible to satisfying all of the constraints specified by $y = Xw$, we need to minimize the following cost function, known as the **least squares objective**

$$\hat{w} = \arg \min_w \frac{1}{2} \|Xw - y\|_2^2$$

The corresponding solution known as **ordinary least squares (OLS)** is obtained using the **left pseudo inverse** or by taking the derivative w.r.t w , $\nabla_w RSS(w) = 0$,

$$X^T(Xw - y) = 0 \rightarrow w_{OLE} = \underbrace{(X^T X)^{-1}}_{\mathbb{R}^{D \times D}} X^T y$$

Machine Learning

└ Linear Regression

└ Least Squares Estimation

└ Least Squares Estimation

What happens if:

we have an additional data point $0 = -w_1 + 3w_2$?

Given the following systems of equations,

$$2 = 3w_1 + 2w_2$$

$$-2 = 2w_1 - 2w_2$$

$$0 = -w_1 + 3w_2$$

it can be written in the matrix form $y = Xw$ as

$$y = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \\ -1 & 3 \end{pmatrix}$$

Since $N > D$, we can simply solve the systems using

$$w_{OLE} = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.18 & 0.48 \end{pmatrix}^T$$

LEAST SQUARES ESTIMATION

When $N > D$ (tall and skinny), the system is overdetermined, so there is no unique solution → Least square estimation?

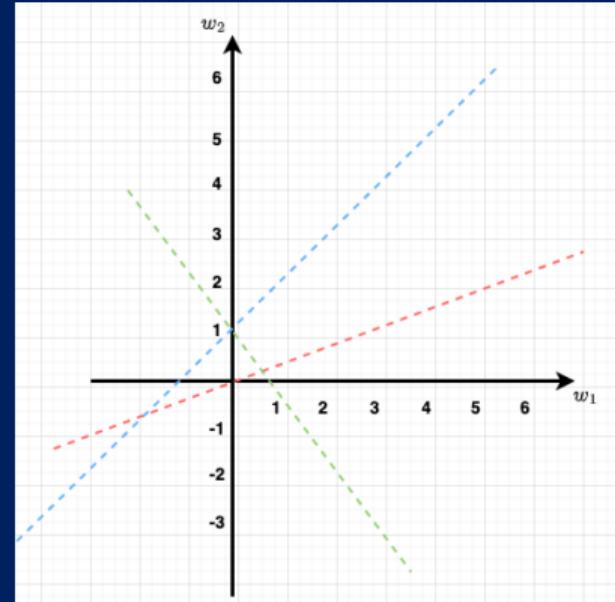
Least Squares Estimation (LSE)

To find the solution that gets as close as possible to satisfying all of the constraints specified by $y = Xw$, we need to minimize the following cost function, known as the least squares objective

$$\hat{w} = \arg \min_w \frac{1}{2} \|Xw - y\|_2^2$$

The corresponding solution known as ordinary least squares (OLS) is obtained using the left pseudo inverse or by taking the derivative w.r.t w , $\nabla_w JLS(w) = 0$,

$$X^T(Xw - y) = 0 \rightarrow w_{OLE} = \underbrace{(X^T X)^{-1}}_{\text{left pseudo inverse}} X^T y$$



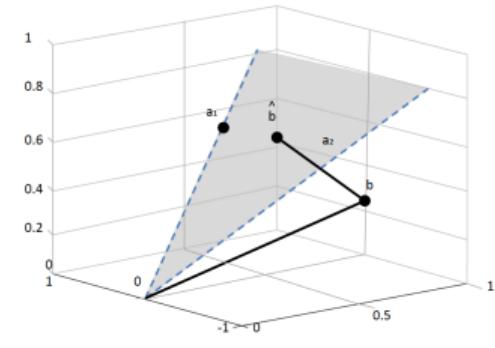
GEOMETRIC INTERPRETATION OF LEAST SQUARES

Our objective is to find the optimal parameters that minimizes the following objective function:

$$\arg \min_{\hat{y} \in \text{span}([x_{:,1}, \dots, x_{:,d}])} \|y - \hat{y}\|_2$$

where $x_{:,d}$ is the d^{th} -column of matrix X and $\hat{y} = Xw$ is the prediction which belongs to the $\text{span}(X)$.

It turns out that the shortest path with minimal distance (residuals) is the **orthogonal projection** of y into the subspace $\text{span}(X)$, i.e., $x_{:,D} \perp (y - \hat{y})$. This is translated to: $x_{:,D}^T(y - \hat{y}) = 0 \rightarrow X^T(y - Xw) \triangleq X^T y - X^T X w = 0$, thus $w_{opt} = (X^T X)^{-1} X^T y \rightarrow$ **ordinary least squares (OLS)**.



²The figure considers $b = Ax$

Machine Learning

└ Linear Regression

└ Least Squares Estimation

└ Geometric Interpretation of least squares

Given the following systems of equations $y = Xw$ as

$$y = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \\ -1 & 3 \end{pmatrix},$$

the geometric interpretation of the residual sum of squares is presented on the right hand side where

$$\begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} w_1 + \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} w_2 = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}$$

$$w_{OLE} = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.18 & 0.48 \end{pmatrix}^T$$

$$\hat{y} = Proj(x)y = \begin{pmatrix} 1.49 & -0.61 & 1.27 \end{pmatrix}^T$$

GEOMETRIC INTERPRETATION OF LEAST SQUARES

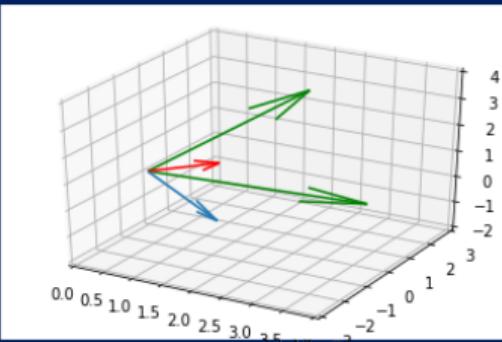
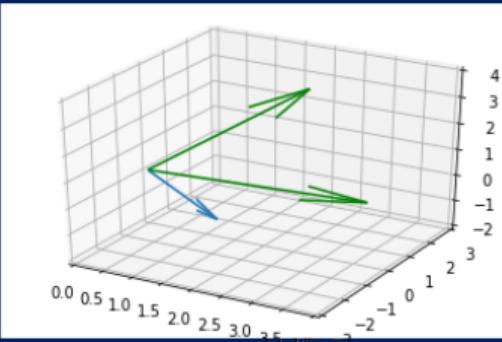
Our objective is to find the optimal parameters that minimizes the following objective function

$$\arg \min_{\hat{y} \in \text{span}\{x_1, \dots, x_d\}} \|y - \hat{y}\|_2$$

where $x_{:,d}$ is the d^{th} -column of matrix X and $\hat{y} = Xw$ is the prediction which belongs to the $\text{span}(X)$.

It turns out that the shortest path with minimal distance (residuals) is the **orthogonal projection** of y into the subspace $\text{span}(X)$, i.e., $x_{:,d} \perp (y - \hat{y})$. This is translated to: $x_{:,d}^T(y - \hat{y}) = 0 \rightarrow X^T(y - Xw) \triangleq X^T y - X^T Xw = 0$, thus $w_{opt} = (X^T X)^{-1} X^T y \rightarrow \text{ordinary least squares (OLS)}$.

²The figure considers $\hat{y} \subset Ax$



ALGORITHMIC ISSUES

When $N \gg D$ (tall and skinny), the system is **overdetermined**, so there is **no unique solution** →

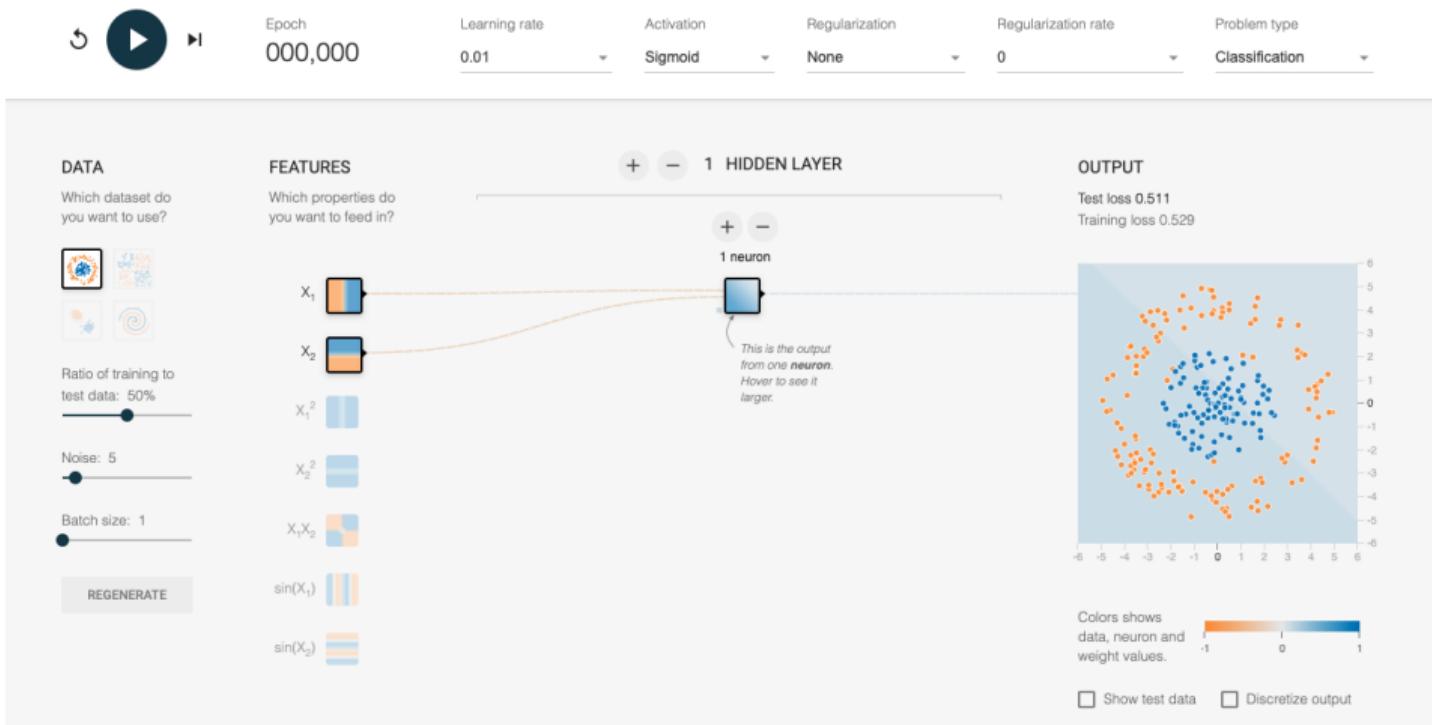
$$w_{OLE} = \underbrace{(X^T X)^{-1}}_{\mathbb{R}^{D \times D}} X^T y$$

numerical reasons – $X^T X$ may be **ill conditioned** or **singular** (look at this example)

alternative and less expensive solutions are **SVD** and **QR** decompositions

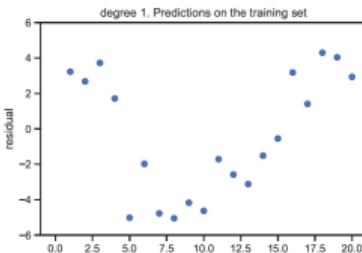
alternative to direct methods based on matrix decomposition is **iterative solvers**
standardize the data (see Sec. 10.2.8)

DEMO

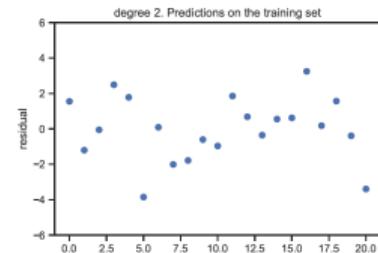


MEASURING GOODNESS OF FIT -- RESIDUAL PLOT

Residual plot for 1D feature

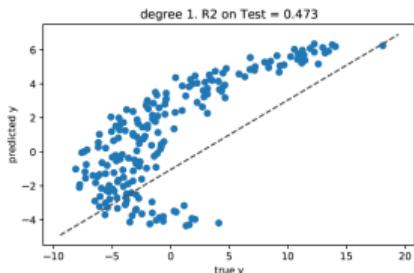


(a)

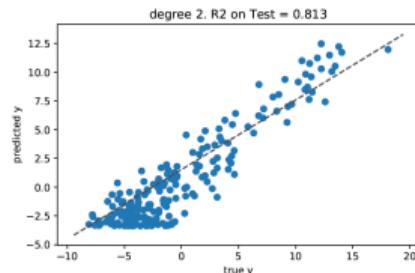


(b)

Residual plot for Multi-dimensional feature



(a)



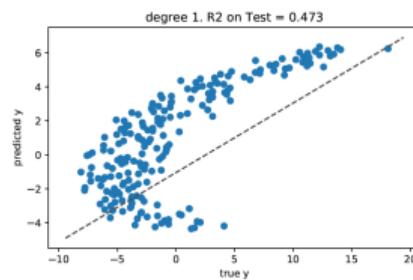
(b)

MEASURING GOODNESS OF FIT -- PREDICTION ACCURACY AND R^2

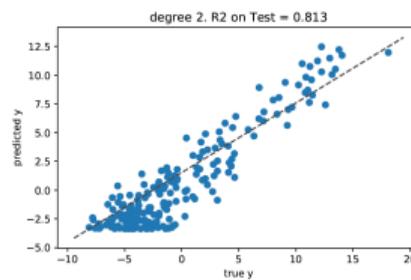
Residual Sum of Square (RSS): $\frac{1}{2} \sum_{n=1}^N (y_n - \hat{y}_n)^2$

Root Mean Square Error (RMSE): $\sqrt{\frac{1}{N} \text{RSS}}$

Coefficient of determination: $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$



(a)



(b)

Questions