1.1 Verify Sort

$$W^{(1)} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$b^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$W^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$b^{(2)} = -2.5$$

1.2 Perform Sort

The modular network in function is previous verify-sort.

Given the input x_1, x_2, x_3, x_4, create $4 \times 3 \times 2 \times 1 = 24$ permutation for $\{x_1, x_2, x_3, x_4\}$ as input layer.

For each permutation, apply verify-sort to it, and output into a node. So in total there are 24 node where we can check which one is correct order using one more network layer to get the final output.

1.3.1

$$W_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad b_0 = \begin{bmatrix} -a \\ b \end{bmatrix}$$

$$W_1 = [h, h] \qquad b_1 = -h$$

1.3.2

$$\|f - \hat{f}_0\| = \int_{-1}^{1} |-x^2+1| - 0 \, dx$$

$$= \int_{-1}^{1} -x^2+1 \, dx$$

$$= \left(-\frac{1}{3}x^3 + x\right)\Big|_{-1}^{1}$$

$$= \left(-\frac{1}{3}+1\right) - \left(\frac{1}{3}-1\right)$$

$$= \frac{2}{3} + \frac{2}{3}$$

$$= \frac{4}{3}$$

$$\|f - \hat{f}_1\| = \int_{-1}^{1} |-x^2+1 - g(h_1, a_1, b_1, x)|\, dx$$

$$= \int_{-1}^{a_1} -x^2+1\, dx + \int_{a_1}^{b_1} -x^2+1-h_1\, dx + \int_{b_1}^{1} -x^2+1\, dx$$

$$= \left(-\frac{1}{3}x^3+x\right)\Big|_{-1}^{a_1} + \left(-\frac{1}{3}x^3+x-h_1x\right)\Big|_{a_1}^{b_1} + \left(-\frac{1}{3}x^3+x\right)\Big|_{b_1}^{1}$$

$$= \left[-\frac{1}{3}a_1^3+a_1-\left(\frac{1}{3}-1\right)\right] + \left|-\frac{1}{3}b_1^3+b_1-bh_1 \quad -\left(-\frac{1}{3}a_1^3+a_1-ah_1\right)\right|$$

$$+\left[-\frac{1}{3}+1-\left(-\frac{1}{3}b_1^3+b_1\right)\right]$$

$$= \left(-\frac{1}{3}a_1^3+a_1+\frac{2}{3}\right) + \left|-\frac{1}{3}b_1^3+b_1+\frac{1}{3}a_1^3-a_1+(a_1-b_1)h_1\right|$$

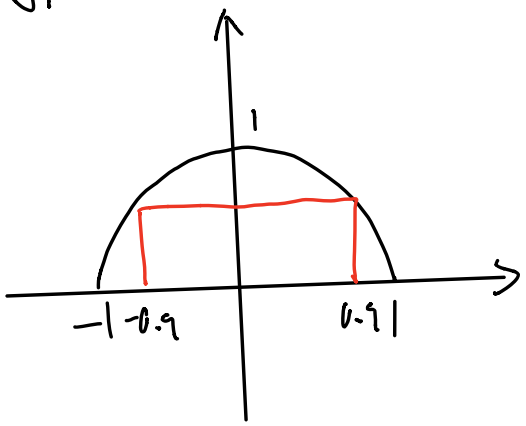
$$+\left(\frac{1}{3}b_1^3-b_1+\frac{2}{3}\right)$$

Want $\|f-\hat{f}_1\| < \frac{4}{3}$

a possible sol:

$a_1 = -0.9 \quad b_1 = 0.9 \quad h_1 = 0.19$

Plot f and $\hat{f}_1$



blade: f

red: $\hat{f}_1$

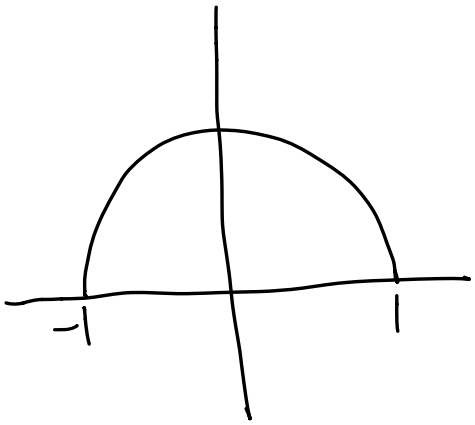
1.3.3 At State $i+1$, $a_{i+1} = a_i + 0.01$

$b_{i+1} = b_i - 0.01$

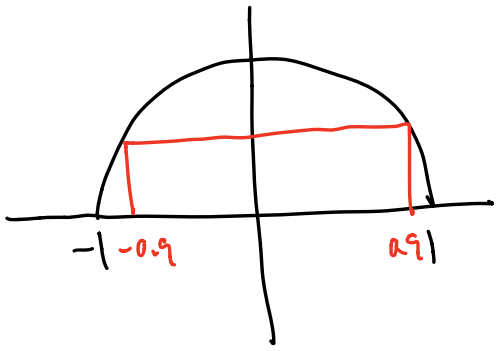
$h_{i+1} = -a_{i+1}^2 + 1$

$\Rightarrow \hat{f}_{i+1}(x) = \hat{f}_i(x) + g(h_{i+1}, a_{i+1}, b_{i+1}, x)$

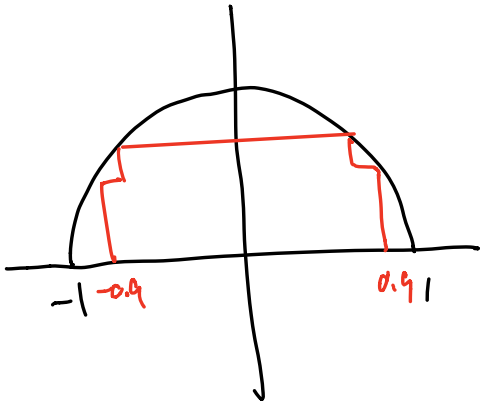
f:



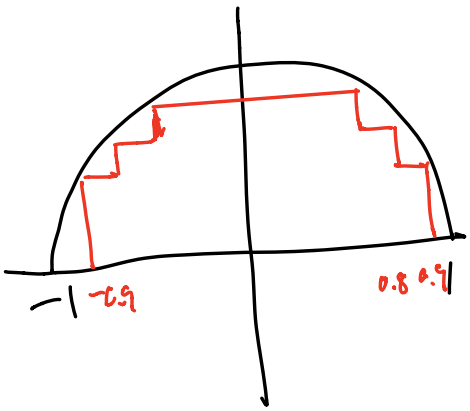
$\hat{f}_1$:



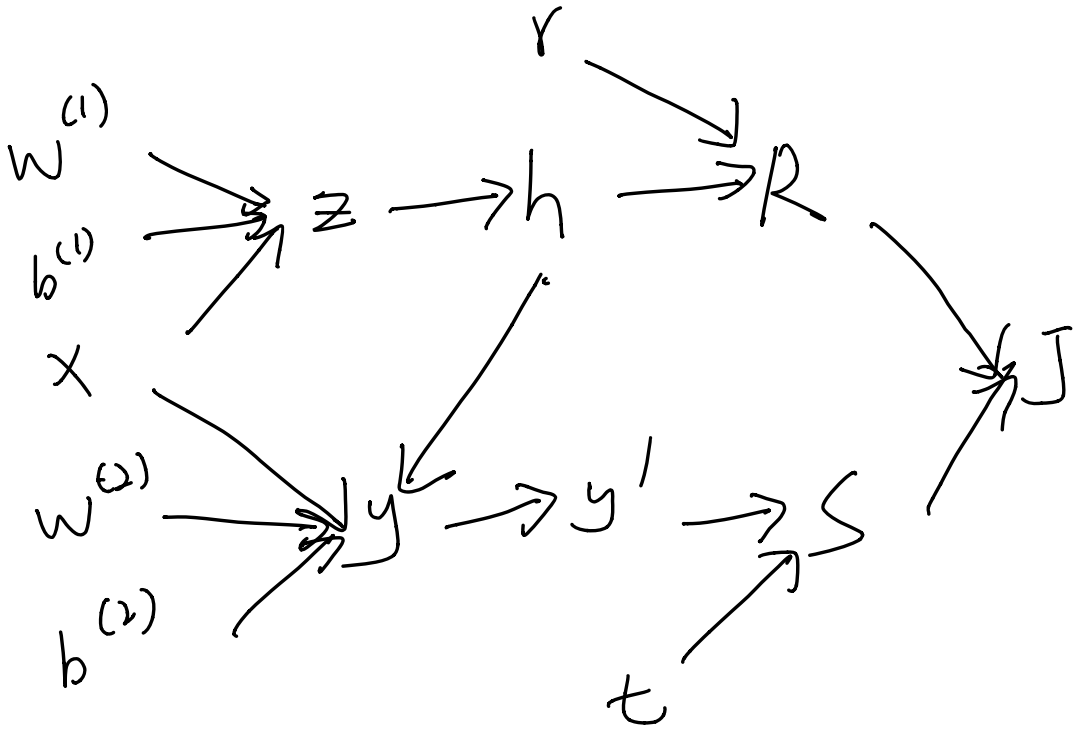
f_2:



f_3



2.1.1



2.1.2

$\bar{J} = 1$

$\bar{R} = \bar{J} = 1$

$\bar{s} = \bar{J} = 1$

$$\bar{y'_k} = \bar{S} \frac{\partial S}{\partial y'_k}$$

$$= \begin{cases} 0 & \text{otherwise} \\ 1 & t = k \end{cases}$$

$$\bar{y} = \bar{y'} \circ \text{softmax}'$$

$$\bar{h} = \bar{y} \cdot \frac{\partial y}{\partial h} + \bar{R} \cdot \frac{\partial R}{\partial h}$$

$$= W^{(2)T} \bar{y} + \gamma$$

$$\bar{z} = \bar{h} \frac{\partial h}{\partial z} = \bar{h} \circ V \quad \text{where}$$

$$V_i = \begin{cases} 0 & z_i \leq 0 \\ 1 & z_i > 0 \end{cases}$$

$$\bar{x} = \bar{z} \frac{\partial z}{\partial x} + \bar{y} \frac{\partial y}{\partial x}$$

$$= W^{(l)T} \bar{z} + \bar{y}$$

2.2.1

$$\frac{\partial}{\partial x} f(x) = v v^T$$

$$= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

2.2.2

time cost $\in O(n^2)$

memory cost $\in O(n^2)$

2.2.3

$$J^T y = \begin{pmatrix} j_{1,1} & j_{2,1} & \cdots & j_{n,1} \\ j_{1,2} & j_{2,2} & & j_{n,2} \\ \vdots & \vdots & \ddots & \\ j_{1,n} & j_{2,n} & & j_{n,n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= V V^T y$$

$$= V (V^T y)$$

calculate $V^T y$ first

then calculate $V (V^T y)$

$$v^T y = (1,2,3)\begin{pmatrix}1\\1\\1\end{pmatrix} = 6$$

$$v(v^T y) = \begin{pmatrix}1\\2\\3\end{pmatrix} \cdot 6 = \begin{pmatrix}6\\12\\18\end{pmatrix}$$

$$\Rightarrow z^T = [6, 12, 18]$$

3.1

$$L = \frac{1}{n}\sum_{i=1}^{n} (\hat{w}^T x^{(i)} - t_i)^T (\hat{w}^T x^{(i)} - t_i)$$

$$= \frac{1}{n}(X\hat{w} - t)^2$$

$$= \frac{1}{n}(X\hat{w} - t)^T (X\hat{w} - t)$$

$$= \frac{1}{n}(\hat{w}^T X^T - t^T)(X\hat{w} - t)$$

$$= \frac{1}{n}(\hat{w}^T X^T X\hat{w} - \hat{w}^T X^T t - t^T X\hat{w} + t^2)$$

$$= \frac{1}{n}(\hat{w}^T X^T X\hat{w} - 2\hat{w}^T X^T t + t^2)$$

So $\frac{\partial L}{\partial \hat{w}} = \frac{1}{n}(2X^T X\hat{w} - 2X^T t)$

3.2.1

$$\frac{\partial L}{\partial \hat{w}} = \frac{1}{n}(2x^T x\hat{w} - 2x^T t) = 0$$

$$\Rightarrow \quad x^T x\hat{w} = x^T t$$

Since $d < n$, we know $x^T x$ is invertible

$$\Rightarrow (x^T x)^{-1} x^T x\hat{w} = (x^T x)^{-1} x^T t$$

$$\Rightarrow \quad \hat{w} = (x^T x)^{-1} x^T t$$

3.2.2

When $d < n$, $\hat{w} = (x^T x)^{-1} x^T t$ is unique since $x^T x$ is invertible.

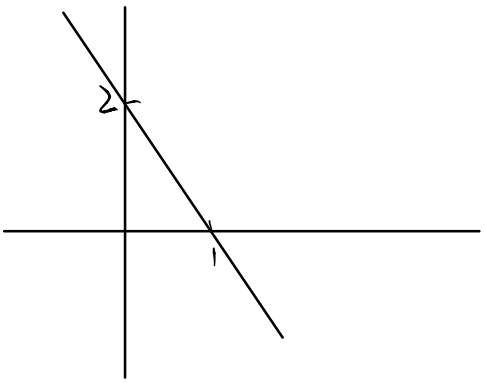
3.3.1

Want to solve $X^T t = X^T X \hat{w}$

$$\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} [2, 1] \hat{w}$$

$$\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \hat{w}$$

$\Rightarrow \hat{w}_2 = -2\hat{w}_1 + 2$



So there exists infinite many solution.

3.3.2

With $n=1$, $\hat{w}(0)=0$

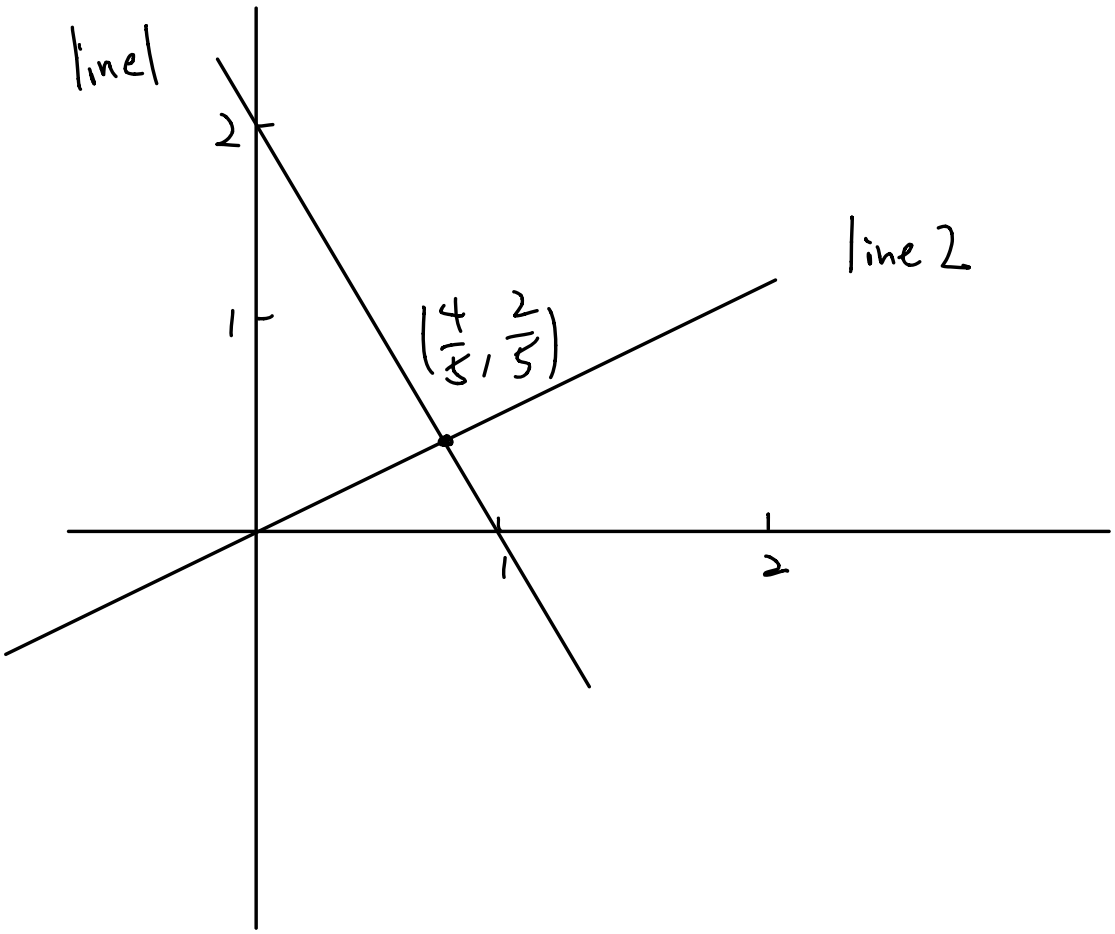
$$\Rightarrow \frac{\partial L}{\partial \hat{w}} = -2X^T t = \begin{bmatrix} -8 \\ -4 \end{bmatrix} \Rightarrow \frac{1}{\sqrt{5}}\begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

which is the direction of gradient.

if we update $\hat{w} = \hat{w}(0) + \alpha \frac{1}{\sqrt{5}}\begin{bmatrix} -2 \\ -1 \end{bmatrix}$

and calculate $\frac{\partial L}{\partial \hat{w}}$ again, then we find new gradient is in the same direction of $\begin{bmatrix} -2 \\ -1 \end{bmatrix}$ which means the gradient never changes along the trajectory.

Thus, it will approach to the minimum lost which is shown in previous question. The gradient solution $\hat{w} = \begin{bmatrix} \frac{4}{5} \\ \frac{2}{5} \end{bmatrix}$


line1
2
1
line 2
$(\frac{4}{5}, \frac{2}{5})$
1
2

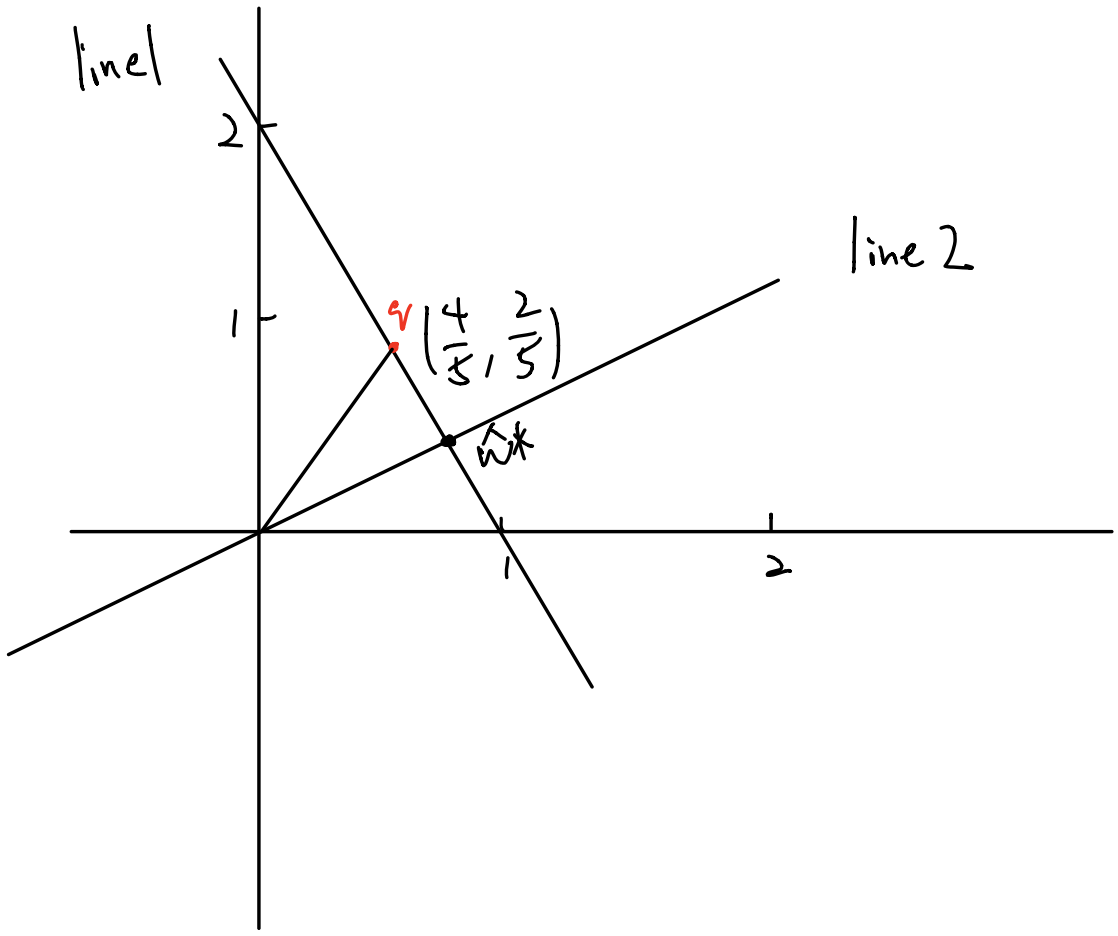
3.3.3

In the following plot, we know line 1 is perpendicular to line1. And the Euclidean norm of vector is the distance between point and origin. So for any point q which is in line 1 not line 2,

we have $\|q\|^2 = \|\hat{w}^*\|^2 + \|\hat{w}^* - q\|^2$

So $\|q\|^2 > \|\hat{w}^*\|^2$

So the gradient solution has smallest Euclidean norm.


line1
2
1
q
$(\frac{4}{5}, \frac{2}{5})$
line 2
$\hat{w}^*$
1
2

3.4.1

We have known following facts about gradient solution:

1. gradient vector is spanned by the rows of X

2. gradient solution has the smallest Euclidean norm

3. The gradient solution is the intersetion between span of rows of X and the space $\hat{w}$ is in.

$\Rightarrow$ the problem is same as

$$\min_{\hat{w}} \hat{w}^2 \quad \text{s.t.} \quad X\hat{w} = t$$

where we can apply Lagrange multiplier.

$$L = \hat{w}^2 - \lambda^T (t - X\hat{w})$$

$$\frac{\partial L}{\partial \hat{w}} = 2\hat{w} + X^T\lambda$$

let $2\hat{w} + X^T\lambda = 0$

$$X^T\lambda = -2\hat{w}$$

$$XX^T\lambda = -2X\hat{w}$$

$$\lambda = -2(XX^T)^{-1}X\hat{w}$$

$$= -2(XX^T)^{-1}t$$

substitute $\lambda = -2(xx^T)^{-1}t$ into

$x^T\lambda = -2\hat{w}$ and get

$$\hat{w} = x^T(xx^T)^{-1}t$$

3.4.2

$\hat{w} = x^T(xx^T)^{-1}t$ from above

wTk $(\hat{w} - \hat{w}_1)\hat{w}$

$$= \left(x^T(xx^T)^{-1}t - \hat{w}_1\right)x^T(xx^T)^{-1}t$$

$$= t^T(xx^T)^{-T}xx^T(xx^T)^{-1}t - \hat{w}_1x^T(xx^T)^{-1}t$$

$$= t^T(xx^T)^{-1}t - \hat{w}_1^Tx^T(xx^T)^{-1}t$$

$$= (t - x\hat{w}_1)^T(xx^T)^{-1}$$

$= 0$ ($\hat{w}_1$ is zero-loss)

It means $\hat{w}$ is perpendicular to vector $u = \hat{w} - \hat{w}_1$ can represents the gradient direction. So that by using Pythagorean Theorem again, we know $\hat{w}$ has the smallest Euclidean norm.

3.5.1

```
def fit_poly(X, d,):
  X_expand = poly_expand(X, d=d, poly_type = poly_type)
  if d > n:
    W = linalg.inv(X_expand.T@X_expand)@X_expand.T@t
  else:
    W = X_expand.T@linalg.inv(X_expand@X_expand.T)@t
  return W
```

Over parameterization doesn't always lead to overfitting. e.g. for 70 polynomial degree, test error is not larger than 10 polynomial degree.