LINEAR REGRESSION REMINDERS

MODEL $f(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x})$

Have code to minimize $\sum_n \left(y^{(n)} - \underline{w}^T \underline{\phi}(\underline{x}^{(n)})\right)^2$ wrt $\underline{w}$

$= (\underline{y} - \underline{\Phi}\underline{w})^T (\underline{y} - \underline{\Phi}\underline{w})$

$\underline{\phi}(\underline{x}) = [\phi_1(\underline{x}) \;\; \phi_2(\underline{x}) \;\; \ldots \;\; \phi_K(\underline{x})]^T$

$\phi_k(\underline{x})$ any scalar function:

- Monomial, eg x_2, $x_3 x_4^3$, ..., 1, ...
- Radial Basis Function

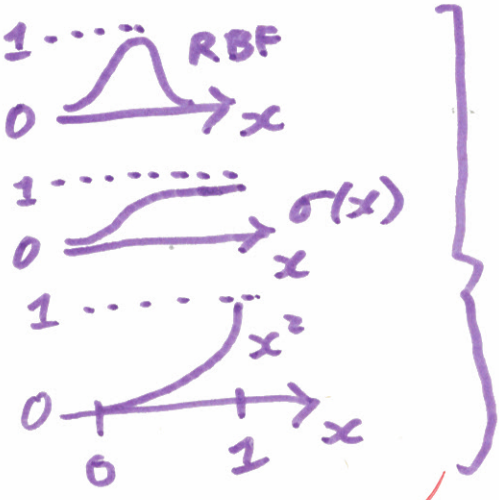
" $\underline{w}^T$  " = 

- Logistic sigmoid 

We noticed:

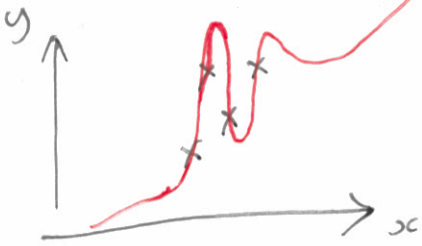
a) Extrapolate differently; b) Placement matters. Might need large K.

Why are large weights bad?



If basis fns are bounded

Large $f \Rightarrow$ large $\underline{w}$

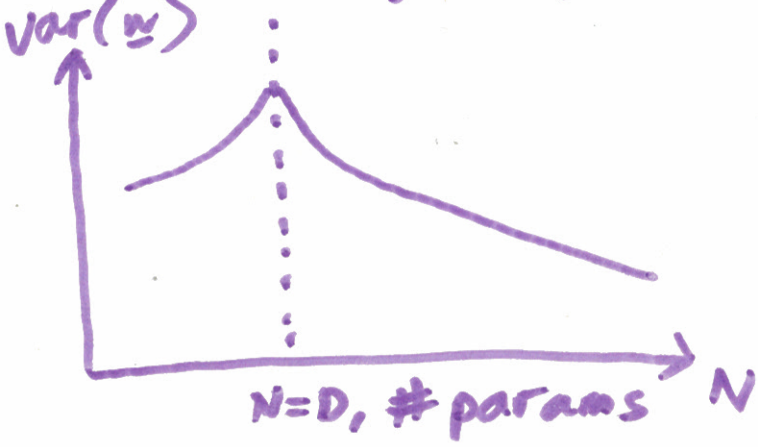


Large derivatives

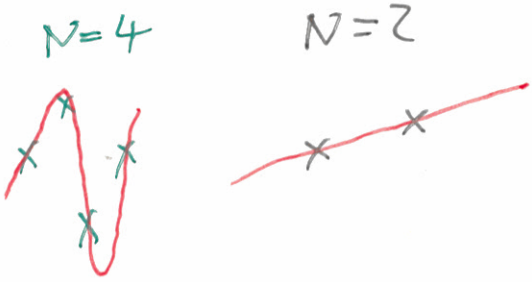
$\Rightarrow$ we've taken extreme differences between ϕ's

$\Rightarrow$ large $\underline{w}^T\underline{w}$

Is amount of overfitting $= \underline{w}^T \underline{w}$?



var: $\frac{1}{K}\sum w_k^2 - \left[\frac{1}{K}\sum w_k\right]^2$



Pseudo-Data Trick

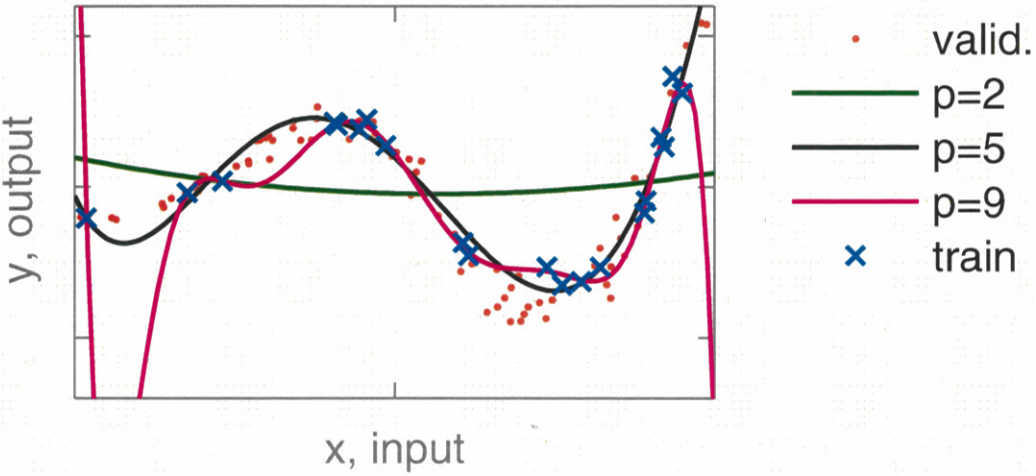
$$\underline{y}' = \begin{bmatrix} \underline{y} \\ \underline{0}_K \end{bmatrix} \Big\} K \text{ fake observations}$$

$$y_{N+1} = 0$$
$$\vdots$$
$$y_{N+K} = 0$$

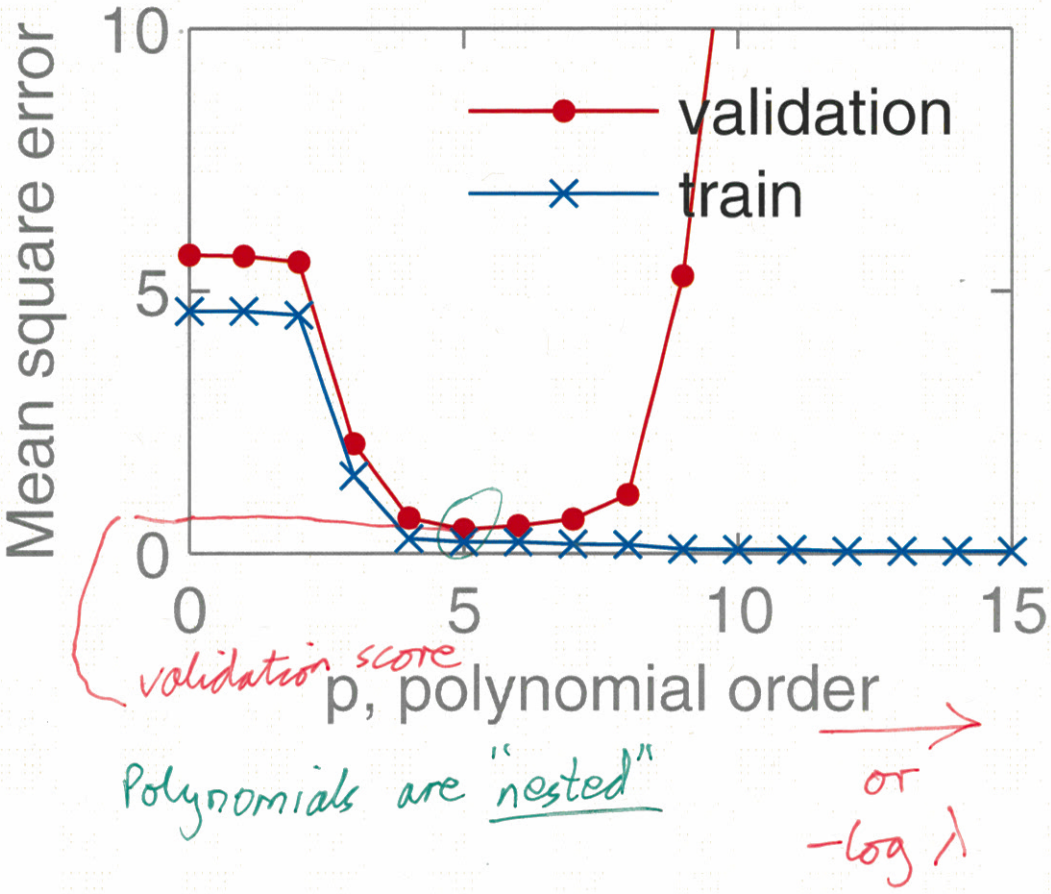
$$\underline{\Phi}' = \begin{bmatrix} \underline{\Phi} \\ \sqrt{\lambda}\, I \end{bmatrix}$$

$$E_\lambda(\underline{w}) = (\underline{y}' - \underline{\Phi}'\underline{w})^T (\underline{y}' - \underline{\Phi}'\underline{w})$$
$$= (\underline{y} - \underline{\Phi}\underline{w})^T (\underline{y} - \underline{\Phi}\underline{w}) + \lambda \underline{w}^T \underline{w}$$
$$(\lambda > 0)$$

Can't use E_λ to set λ


y, output
x, input
valid.
p=2
p=5
p=9
train

should be a log scale



validation score

Polynomials are "nested"

or $-\log \lambda$

TRAIN ERROR MONOTONIC BECAUSE MODELS ARE NESTED

Generalisation

$$\mathbb{E}_{p(x,y)}\left[L(y, f(x))\right] = E_{gen}$$

Loss, e.g. $(y - f(x))^2$

We assume there is some fixed distribution on future inputs + outputs

$$E_{gen} = \int L(y, f(\underline{x}))\, p(\underline{x}, y)\, d\underline{x}\, dy$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} L(y^{(m)}, f(\underline{x}^{(m)})) = E_{test}$$

$\underline{x}^{(m)}, y^{(m)} \sim p(\underline{x}, y)$

Drawn from a Test set

Unbiased:

$$\mathbb{E}[E_{test}] = \frac{1}{M} \sum_{m=1}^{M} \underbrace{\mathbb{E}[L_m]}_{E_{gen}}$$

$$= \frac{1}{M} M E_{gen} = E_{gen}$$

But not if model was selected E_{test} small!

Data splits

Training set: fit $\underline{w}$, parameters like those

Validation set: used to pick models
eg. polynomial order
or regularizers λ

Test set: to report how well we do in future.