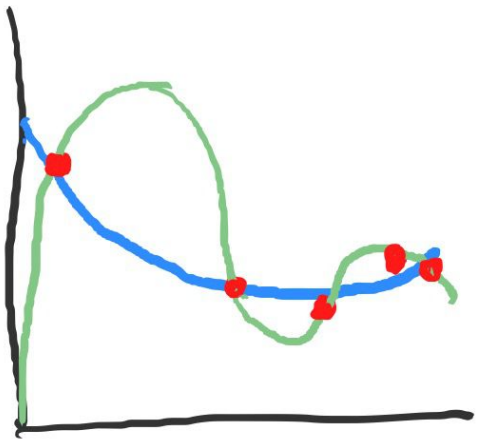
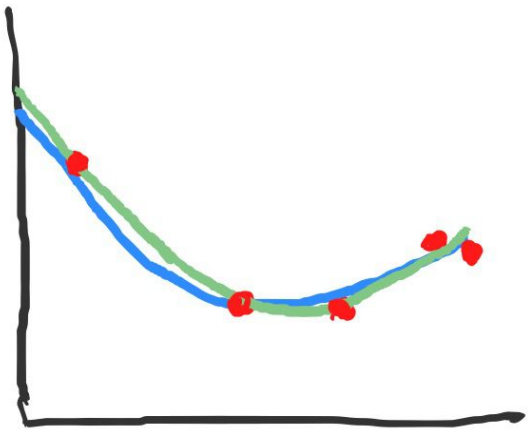
Regularization

Overcoming overfit



(free fit situation)
(No constraint)
(4th order poly)



(Restricted fit)
(4th order poly, but with a regularization term)

Regularization : It constraints the learning algorithm to improve out-sample error when noise is present or we have a lack of training samples.

Regularization from VC analysis :

$$E_{out}(h) \leq E_{in}(h) + \underbrace{\Omega(H)}_{\text{model complexity penalty}} \quad \forall h \in H$$

(depends on H, N, δ)

Idea : · Fit using a 'simple' h from $\mathcal{H}$

So effectively minimizing

$$\underset{h}{\text{minimize}} \quad E_{in}(h) + \Omega(h)$$

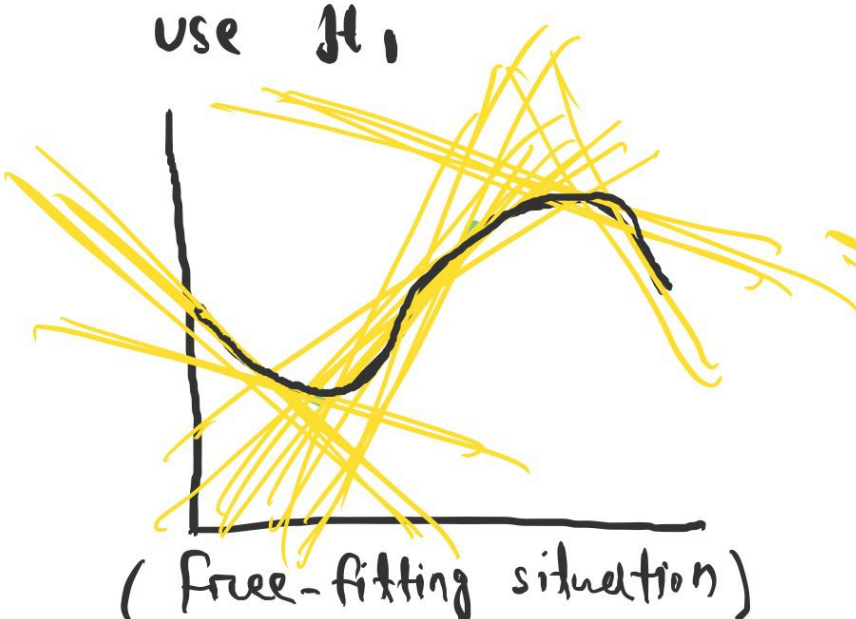
Some kind of measure on h. Ω is something that measure the complexity of h.

Motivation example :

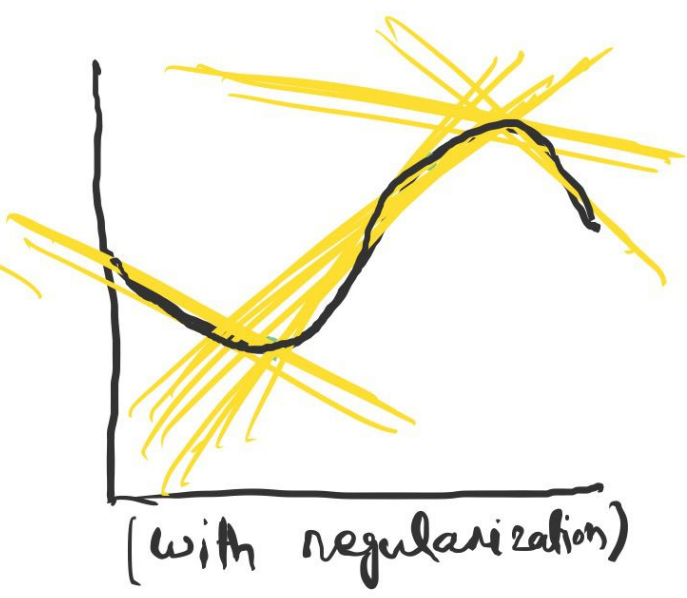
Weight decay technique :

→ measures the complexity of a hypothesis h by the size of the co-efficients used to represent h (e.g. linear model)

→ Apply this to sine example, for $N = 2$. use $\mathcal{H}_1$

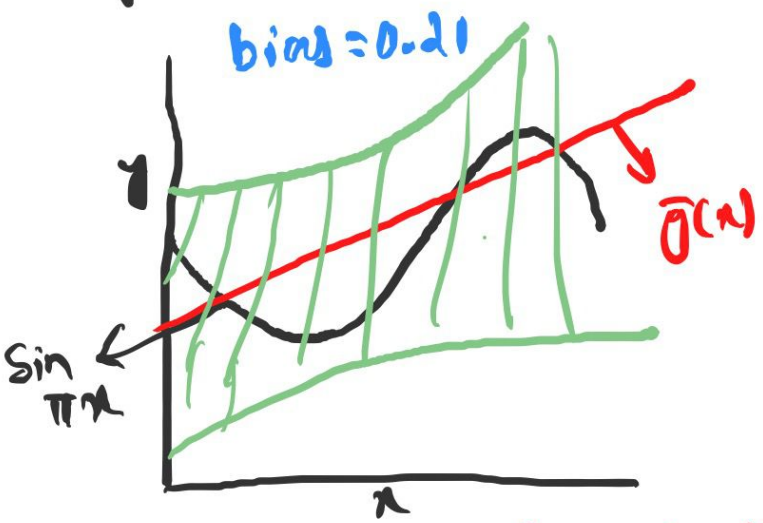


(Free-fitting situation)

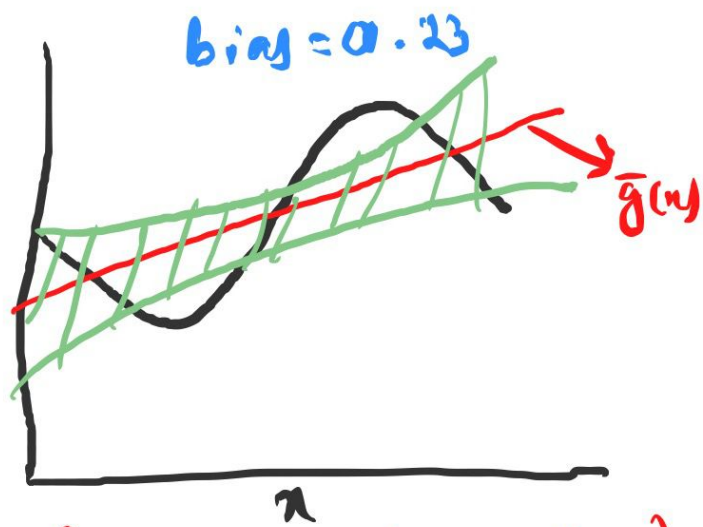


(with regularization)

This weight decay technique reduces the variation by limiting the magnitude of your co-efficients of the linear model.



(without regularization)



(with regularization)

Constant model $E_{out} = 0.75$

Unregularized model $E_{out} = 1.90$

Regularized model $E_{out} = 0.56$

→ Bias-Variance point of view: Improve variance but suffer from bias. Overall is better.

bias: quality of the average fit

Q: Why need regularization?

→ The linear model is too sophisticated for the amount of data we have. Since a line can fit any 2 points.

→ The need of regularization depends on quantity and quality of data.

→ Given any two points, we can either choose
- a simple mode e.g. constant model
- to constrain the model (e.g. weight decay method)

→ Constraining the model gives us more flexibility.

Regularization techniques :

1) Weight decay method (constraint minimization)
2) Augmentation error (uncontraint)

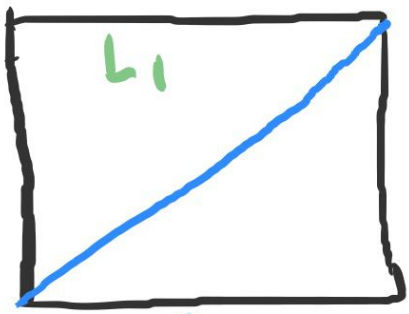
1) Weight decay method (soft order constraint)

→ Consider the following example

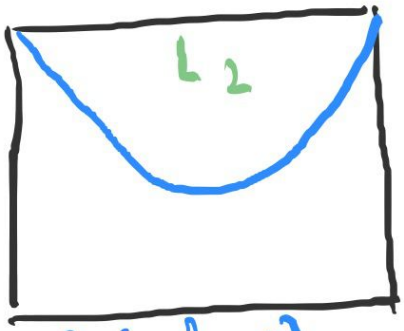
H = set of polynomials in one variable

$x \in [-1, 1]$ e.g. $h(x) = 2x^2 + 3x + 7$

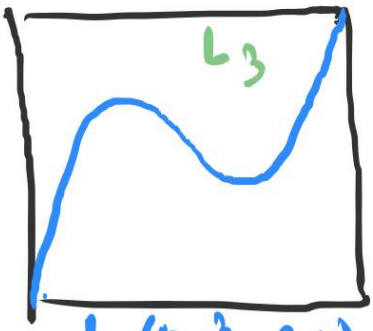
→ Express $h(x)$ using basis function. So the basis functions for polynomials are Legendre polynomials $L_q(x)$, $q = 1, 2, \ldots$



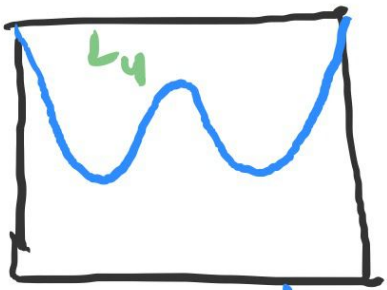
x



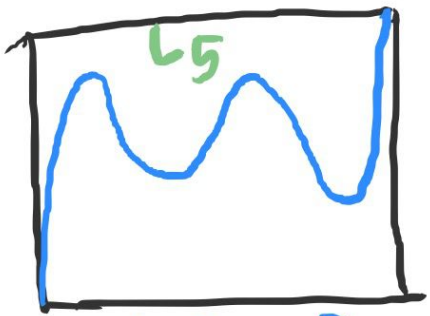
$\frac{1}{2}(3x^2-1)$



$\frac{1}{2}(5x^3-2x)$



$\frac{1}{8}(35x^4-30x^2+3)$



$\frac{1}{8}(63x^5+\dots)$

These polynomials are orthogonal.

So $h(x) = \sum_{q=1}^{Q} w_q L_q(x)$

This model is linear.

If we define a non-linear transform Φ

$$z = \Phi(x) = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix}$$

Then the hypothesis set is

$$\mathcal{H}_Q = \left\{ h \mid h(x) = w^T z = \sum_{q=0}^{Q} w_q L_q(x) \right\}.$$

Now we can define training error (for linear regression)

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T z_n - y_n)^2 .$$

There are multiple ways of constraining the weights.

1) Hard constraint :

→ Force co-efficients to be zero.

→ e.g. $\mathcal{H}_2 = \{ w \mid w \in \mathcal{H}_{10} : w_q = 0, \text{ for } q \geq 3 \}$

2) Soft constraint :

→ Force co-efficients to be small.

→ e.g. $\sum_{q=0}^{Q} w_q^2 \leq C$

→ This encourages weights to be small without changing the order of the polynomial by explicitly forcing some weights to 0.

VC perecpective of soft order constraint:

The optimization problem is

$$\min_{w} E_{in}(w) \quad \text{Subject to} \quad w^T w \leq C.$$

We know $E_{in}(w) = \frac{1}{N} \| Zw - y \|^2$

Z: transformed input vector put into a matrix

y: labels

The soft order constrained hypothesis is

$$\mathcal{H}(C) = \{ h \mid h(x) = w^T z, \; w^T w \leq C \}$$

The optimization is equivalent to saying minimize

E_{in} over $\mathcal{H}(C)$.

→ If $C_1 < C_2$, then $\mathcal{H}(C_1) \subset \mathcal{H}(C_2)$.

and $d_{VC}(\mathcal{H}(C_1)) \leq d_{VC}(\mathcal{H}(C_2))$

→ We should expect better generalization with $\mathcal{H}(C_1)$.

Augmented error (Unconstrained minimization)

$$E_{aug}(w) = E_{in}(w) + \lambda w^T w$$

$\lambda > 0$ is a free parameter.

→ Unconstraine minimization is often easier than constrained minimization.

→ For a given C, soft order constraint corresponds to selecting a hypothesis from a smaller set $\mathcal{H}(C)$.

→ For augemented error, you need to find the optimal parameter λ^*.

VC perspective of augmented error:

The augmented error for a hypothesis $h \in \mathcal{H}$ is

$$E_{aug}(h, \lambda, \Omega) = E_{in}(h) + \frac{\lambda}{N}\Omega(h).$$

Here, $\Omega(h) = w^T w$.

There are two components of the penalty:

1) The regularizer $\Omega(h)$ which penalizes a particular property of h.
2) Regularization parameter λ (controls the amount of regularization)