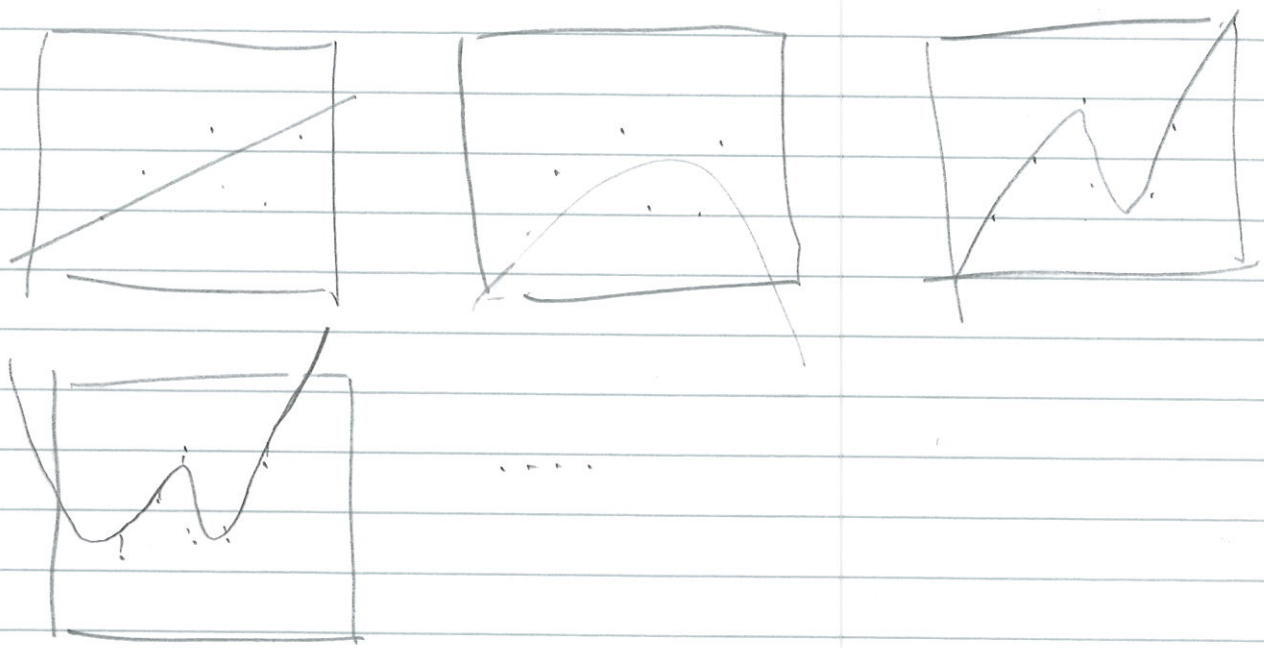
Regularization & Bayesian Linear Regression (1)

As model complexity increases the ability for a model to fit the data becomes higher.

Polynomial regression serves as an excellent pedogogical tool for explaining what happens when a model begins to overfit:



As the polynomial order increases the ability of the model to exactly reproduce the target values increases; however, out-of-sample performance typically degrades.

To combat this we typically regularize or bias models towards reasonable solutions. In this case we might like the polynomial weights to be small.

One way to do this is to directly penalize large weights, ie. modifying our ML objective

$$E_D(\vec{w}) = \frac{1}{2} \sum_{n=1}^{N} \{ t_n - \vec{w}^T \phi(x_n) \}^2$$

$$= \frac{1}{2} \left(\vec{t} - \underline{\Phi} \vec{w} \right)^T \left(\vec{t} - \underline{\Phi} \vec{w} \right)$$

where E is "energy", ie. log. prob. and

$$\underline{\Phi} = \begin{matrix} \\ n \end{matrix} \overset{k}{\begin{bmatrix} \phi_1(x_1) & \cdots & \phi_k(x_1) \\ \vdots & & \vdots \\ \phi_1(x_n) & \cdots & \phi_k(x_n) \end{bmatrix}}$$

to include a term like

$$E_w(\vec{w}) = \frac{\lambda}{2} \vec{w}^T \vec{w}$$

where λ is a smoothing tuning parameter. Maximum probability is minimum energy so we might wish to find

$$\vec{w}^* = \underset{\vec{w}}{\operatorname{argmin}} \; E_D(\vec{w}) + E_w(\vec{w})$$

$$= \underset{\vec{w}}{\operatorname{argmin}} \; \frac{1}{2} \left(\vec{t} - \underline{\Phi} \vec{w} \right)^T \left(\vec{t} - \underline{\Phi} \vec{w} \right) + \frac{\lambda}{2} \vec{w}^T \vec{w}$$

which arises at $\frac{\partial}{\partial \vec{w}} \left(E_D(\vec{w}) + E_w(\vec{w}) \right) = 0$

$$0 = \frac{\partial}{\partial \vec{w}} \left(\frac{1}{2} \left(\vec{t} - \underline{\Phi} \vec{w} \right)^T \left(\vec{t} - \underline{\Phi} \vec{w} \right) + \frac{\lambda}{2} \vec{w}^T \vec{w} \right)$$

$$- \underline{\Phi}^T \left(\vec{t} - \underline{\Phi} \vec{w} \right) + \lambda \vec{w}$$

$$= - \underline{\Phi}^T \vec{t} + \left(\underline{\Phi}^T \underline{\Phi} + \lambda \underline{I} \right) \vec{w}$$

$$\Rightarrow \vec{w} = \left(\underline{\Phi}^T \underline{\Phi} + \lambda \underline{I} \right)^{-1} \underline{\Phi}^T \vec{t}$$

What does this mean and do?

1) Inverting $\Phi^T\Phi$ becomes unstable or impossible if Φ is rank deficient. Adding positive elements to the diagonal ensures that $(\Phi^T\Phi + \lambda I)$ is full rank and therefore invertible.

2) $\frac{\lambda}{2}\vec{w}^T\vec{w}$ looks a lot like a mumormal, ie. $\frac{1}{2}(\vec{w}-\vec{0})^T \lambda I (\vec{w}-\vec{0})$ which in energy terms will be minimized when $\vec{w}$ is close to $\vec{0}$.

3) This is a form of BIAS that is helpful!

To interpret this further we appeal to Bayesian reasoning, e.g. Bayesian linear regression.

Let's consider Bayes rule in the context of linear regression.

$$p(\vec{w} \mid \vec{t}, \Phi, \beta, \lambda) = \frac{p(\vec{t} \mid \Phi, \vec{w}, \beta)\, p(\vec{w} \mid \lambda)}{p(\vec{t} \mid \Phi, \beta, \lambda)}$$

this is a little confusing so let's drop all the constants

$$p(\vec{w} \mid \vec{t}) = \frac{p(\vec{t} \mid \vec{w})\, p(\vec{w})}{p(\vec{t})}$$

$$\propto p(\vec{t} \mid \vec{w})\, p(\vec{w})$$

$$= \mathcal{N}(\vec{t};\ \Phi\vec{w}, \beta I)\, \mathcal{N}(\vec{w};\ \vec{0}, \lambda I)$$