

CS-568 Deep Learning

Regularization in Neural Networks

Nazar Khan

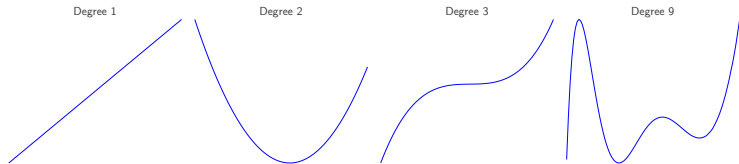
Department of Computer Science

University of the Punjab

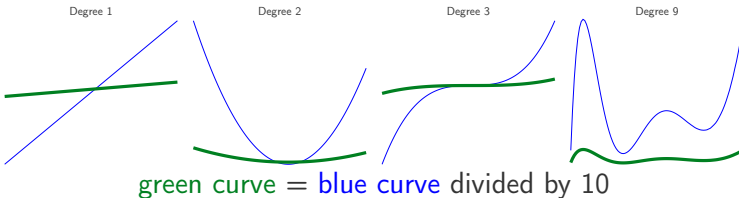
Before we start

A primer on ML

1. Capabilities of polynomials (lines, quadratics, cubics, \dots , degree M).



2. Capability can be reduced by restricting coefficients.



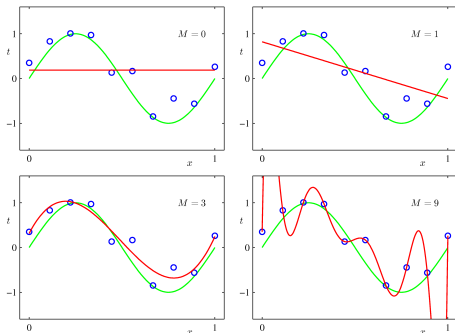
Before we start

A primer on ML

3. Everything is noisy.

$$\text{Observation} = \text{Reality} + \text{Noise}$$

4. Therefore, zero *training* error is bad. Over-fitting vs generalisation.



5. Over-fitting can be reduced via regularization.

Weight Penalties

- ▶ Similar to polynomials, networks with large weights are more powerful.
- ▶ Therefore, more prone to overfitting.
- ▶ So penalise magnitudes of weights to restrict capability.

$$\tilde{L}(\mathbf{w}) = L(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

- ▶ *Hyperparameter*¹ λ controls the level of overfitting.
- ▶ Alternative: separately penalise each layer

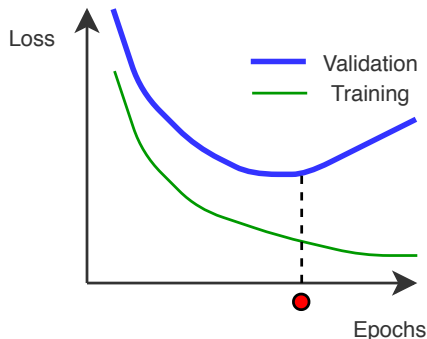
$$\tilde{L}(\mathbf{w}) = L(\mathbf{w}) + \sum_{l=1}^L \frac{\lambda_l}{2} \|\mathbf{w}^{(l)}\|^2$$

Not used often due to increased number of hyperparameters.

¹Something that is not a parameter but influences what the parameters will be.

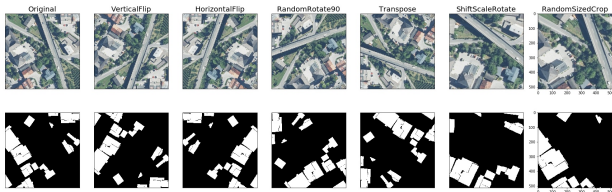
Early Stopping

- ▶ Split some part of the training set into a validation set that will not be used for training.
- ▶ During training, record loss on training as well as validation set.
- ▶ When validation loss starts increasing while training loss is still going down, the model has started overfitting.
- ▶ So stop training at that point.



Data Augmentation

- ▶ Augment training set with transformed versions of training samples.
- ▶ Domain specific data augmentations
 - ▶ Images: Color, Geometry
 - ▶ Text: Synonyms, Tense, Order
 - ▶ Speech: Speed, Sound effects



<https://github.com/albumentations-team/albumentations>

Data Augmentation



<https://github.com/aleju/imgaug>

Label Smoothing

- ▶ Training adjusts the model to make outputs as close as possible to the targets/labels.
- ▶ So if labels are smoothed a little, overfitting will be reduced.
- ▶ For example, if label 0 is mapped to 0.1 and 1 is mapped to 0.9, training will converge early.
- ▶ Training procedure will not try as hard as before to output as close as possible to 0 or 1.

Summary

- ▶ All data contains noise.
- ▶ Given enough power, a neural network will model noise as well.
- ▶ Restricting the network's power allows it to model the underlying behaviour of data instead of noise.
- ▶ This reduces over-fitting on training data and improves generalisation of the network on unseen data.