Lecture 7 (Gradient Descent cotd)

1 When to Stop the Descent

1. $|\theta^{t+1} - \theta^t| \leq \epsilon$
2. $||\nabla_\theta f(\theta)||_1 \leq \epsilon'$ (1-norm), $\epsilon = \epsilon' \cdot \eta$
3. $|f(\theta^{t+1}) - f(\theta^t)| \leq \delta$
4. (alternative) *epoch*, i.e., max value of t

When doing NN, it will be possible to play with these three.

Note: There exists a variation called **Stochastic Gradient Descent (SGD)** too. It works with subset of the examples (training set). This causes the learning curve to not look monotonic *initially*.

2 Choosing Right η

1. When too large - the value of θ^t will oscillate or diverge
2. When too small - the convergence is very slow

3 Concept of Validation Set

This is used to fix the problem of overfitting. The data is divided into $4:1$ training : validation set wherein the error on the validation set is observed. Once this error starts increasing, the descent is stopped. Cross-validation is also possible wherein the data is divided into 5 parts and 5 different models are trained keeping each part as the validation set for every model.