NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 42: Popular CNN Models VI

CONCEPTS COVERED

Concepts Covered:

- CNN
  - Challenges in Deep Learning
  - GoogLeNet
  - ResNet
  - Momentum Optimizer

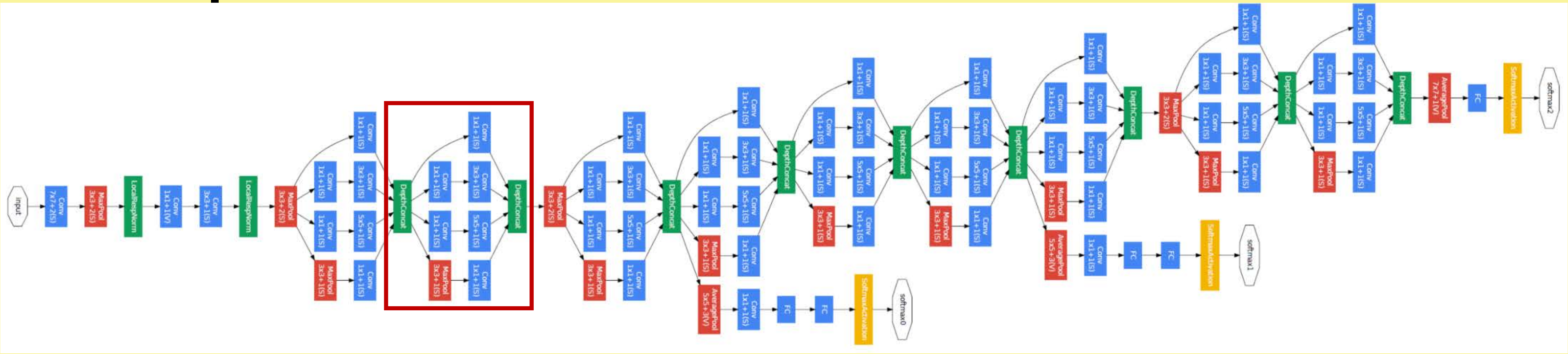
Challenges

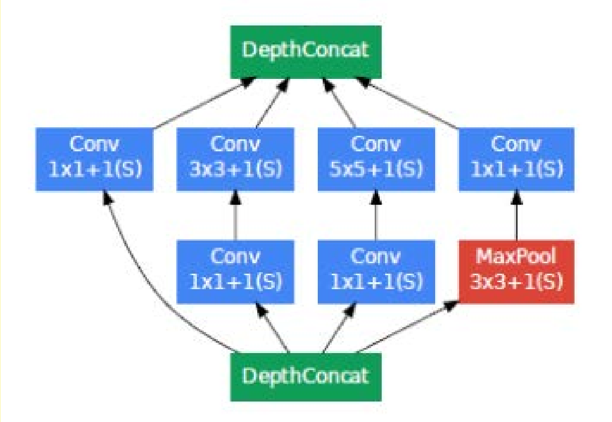
- Deep learning is data hungry.
- Overfitting or lack of generalization.
- Vanishing/Exploding Gradient Problem.
- Appropriate Learning Rate.
- Covariate Shift.
- Effective training.

Vanishing Gradient Problem

- Choice of activation function: ReLU instead of Sigmoid.
- Appropriate initialization of weights.
- Intelligent Back Propagation Learning Algorithm.

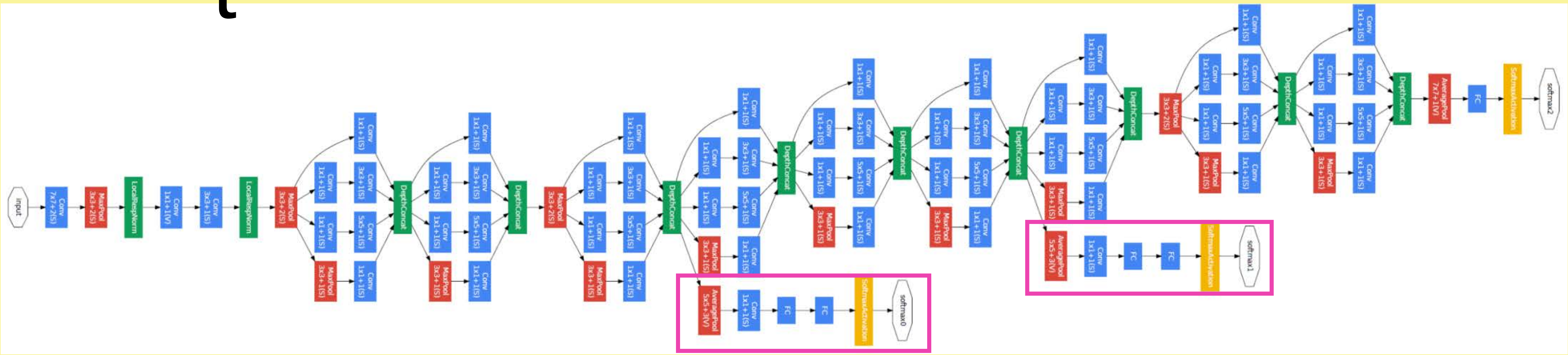
GoogLeNe t

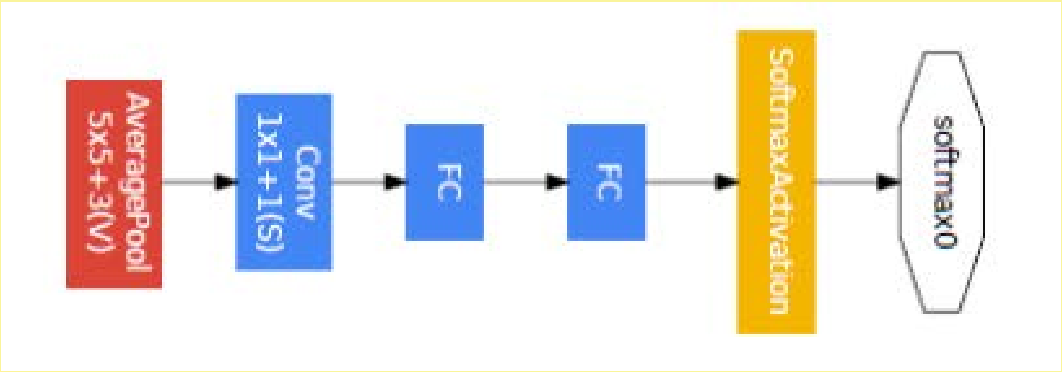




Inception Module

GoogLeNet



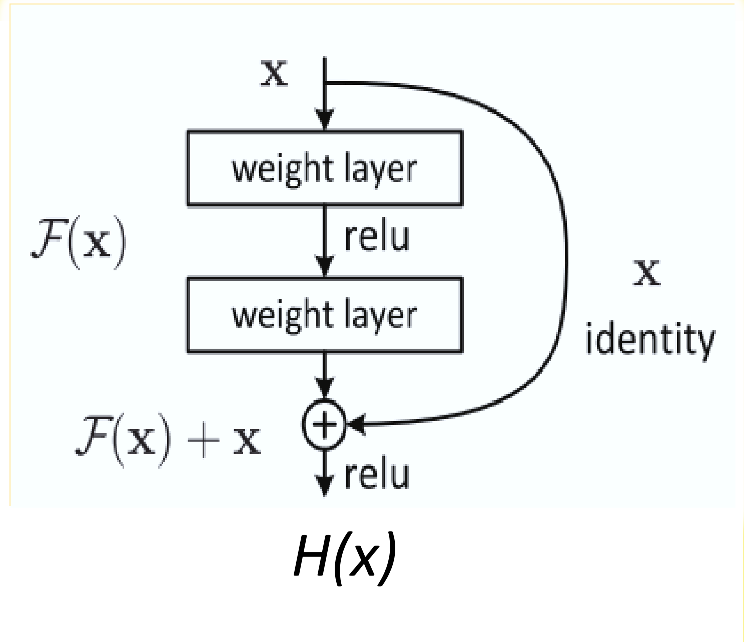


Auxiliary Classifier

ResNet

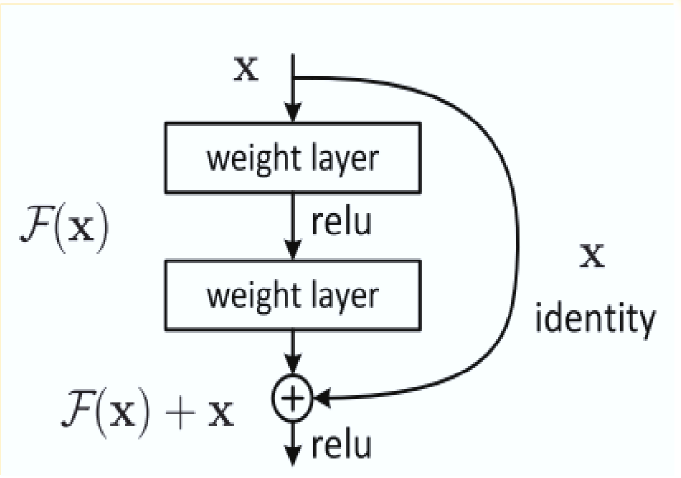
ResNet

- Core idea is: introduction of Skip Connection/ Identity Shortcut Connection that skips one or more layers.
- Stacking layers should not degrade performance compared to its shallow counterpart.
- Weight layer learns *F(x)=H(x)-x*



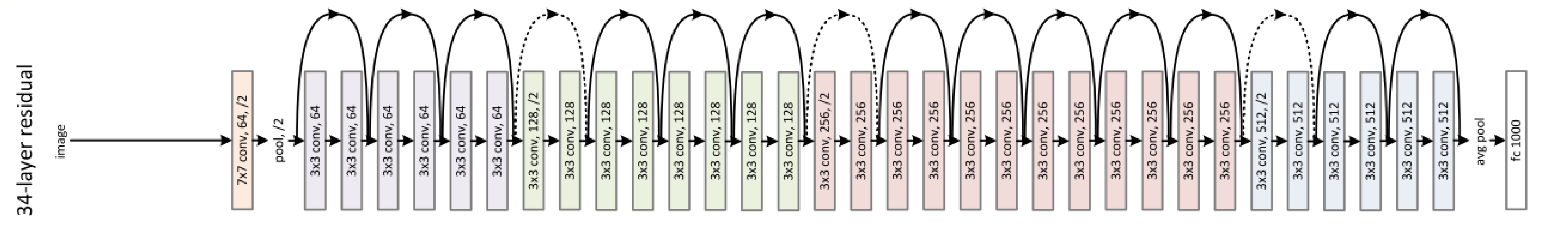
ResNet

- By stacking identity mappings the resultant deep network should give at least same performance as its shallow counterpart.
- Deeper network should not give higher training error than shallow network.
- During learning the gradient can flow to any earlier network through shortcut connections alleviating vanishing gradient problem.



H(x)

ResNet



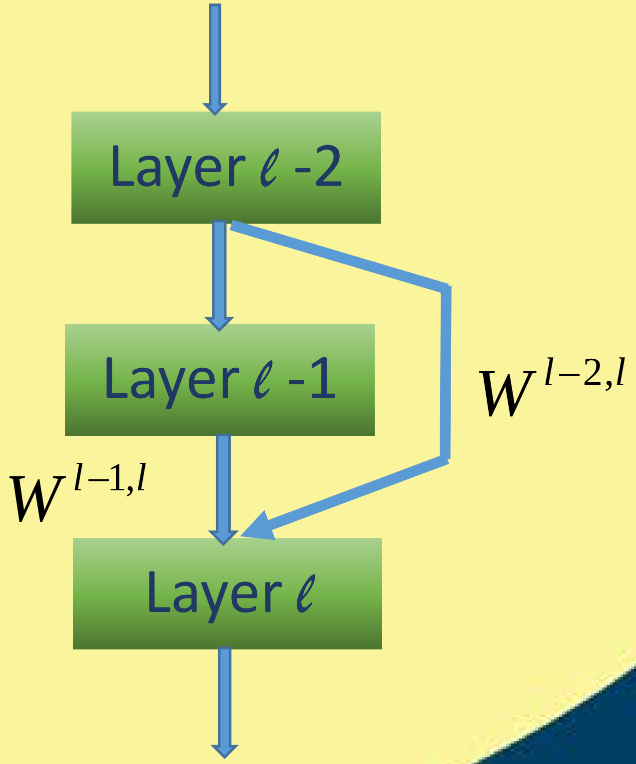
https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

ResNet

Forward flow:

$$a^{l} = f(W^{l-1,l}.a^{l-1} + b^{l} + W^{l-2,l}.a^{l-2})$$

$$= f(Z^{l} + W^{l-2,l}.a^{l-2})$$



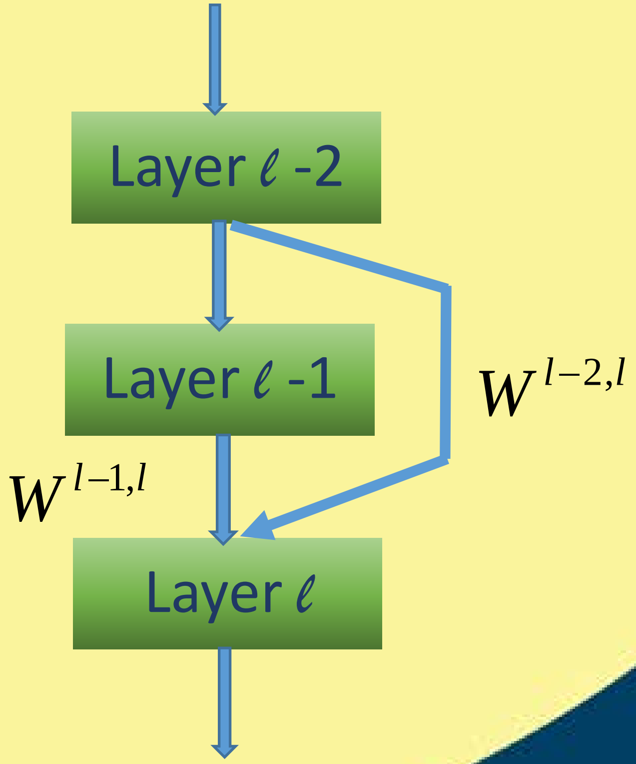
$$a^{l} = f(Z^{l} + a^{l-2}) \qquad \text{if same dimension}$$

ResNet

Backward Propagation:

$$\nabla W^{l-1,l} = -a^{l-1}.\delta^{l} \quad \text{normal path}$$

$$\nabla W^{l-2,l} = -a^{l-2}.\delta^{l} \quad \text{skip path}$$



If the skip path has fixed weights, identity matrix, then they are not updated.

Challenges

- ❑ Deep learning is data hungry.
- ❑ Overfitting or lack of generalization.
- ❑ Vanishing/Exploding Gradient Problem.
- ❑ Appropriate Learning Rate.
- ❑ Covariate Shift.
- ❑ Effective training.

Optimizing
Gradient Descent

Gradient Descent Challenges

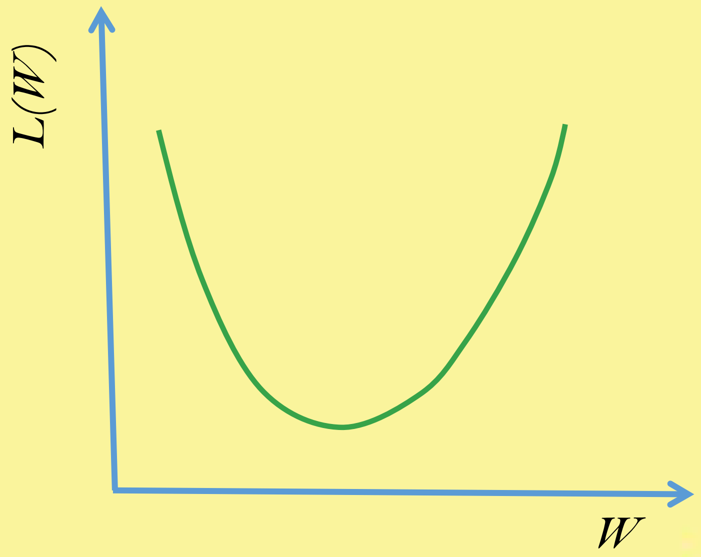
Challenges of Mini-batch Gradient Descent

- Choice of Proper Learning Rate:
  - Too small a learning rate leads to slow convergence.
  - A large learning rate may lead to oscillation around the minima or may even diverge.

Gradient Descent Challenges

Challenges of Mini-batch Gradient Descent

- Choice of Proper Learning Rate:
  - Too small a learning rate leads to slow convergence.
  - A large learning rate may lead to oscillation around the minima or may even diverge.

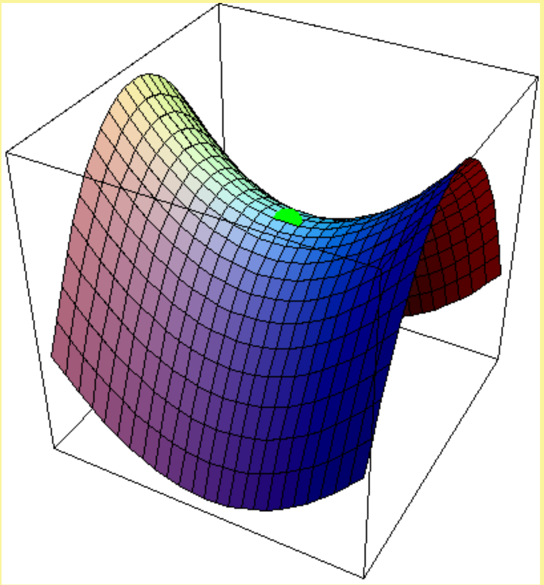


Gradient Descent Challenges

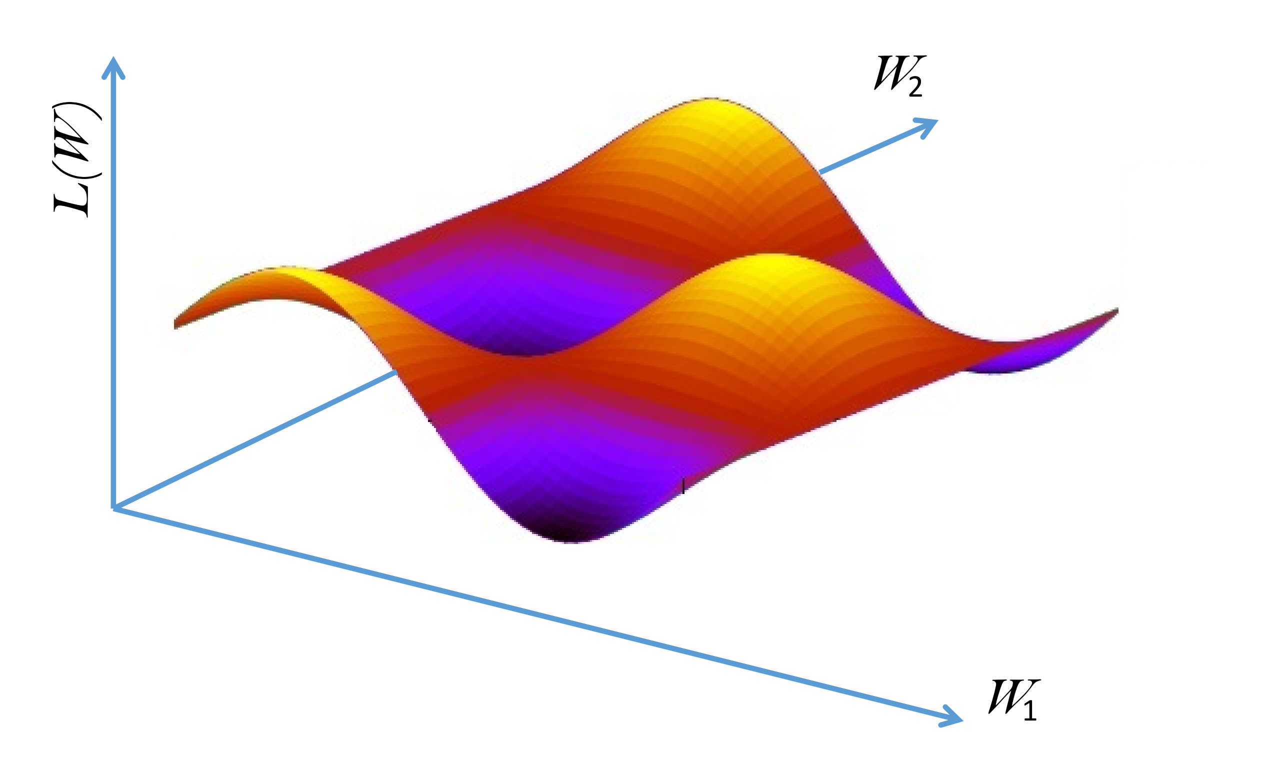
- Learning Rate Schedules: changing learning rate according to some predefined schedule.
  - The same learning rate applies to all parameter updates.
  - The data may be sparse and different features have very different frequencies.
  - Updating all of them to the same extent might not be proper.
  - Larger update for rarely occurring features might be a better choice.

Gradient Descent Challenges

- Avoiding getting trapped in suboptimal local minima.
  - Difficulty arises in from saddle points, i.e. points where one dimension slopes up and another slopes down.
  - These saddle points are usually surrounded by a plateau of the same error, which makes it hard for SGD to escape, as the gradient is close to zero in all dimensions.



Momentum Optimizer



NPTEL ONLINE CERTIFICATION COURSES

Thank you