NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

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

Topic

Lecture 43: Popular Optimizing Gradient Descent

Challenges

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

CONCEPTS COVERED

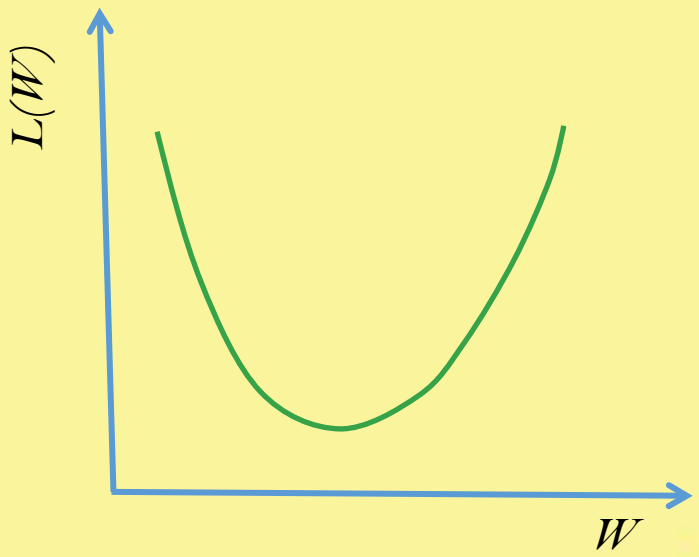
Concepts Covered:

- CNN
  - ResNet
  - Gradient Descent Challenges
  - Momentum Optimizer
  - Nestevor Accelerated Gradient
  - Adagrad.
  - etc.

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.

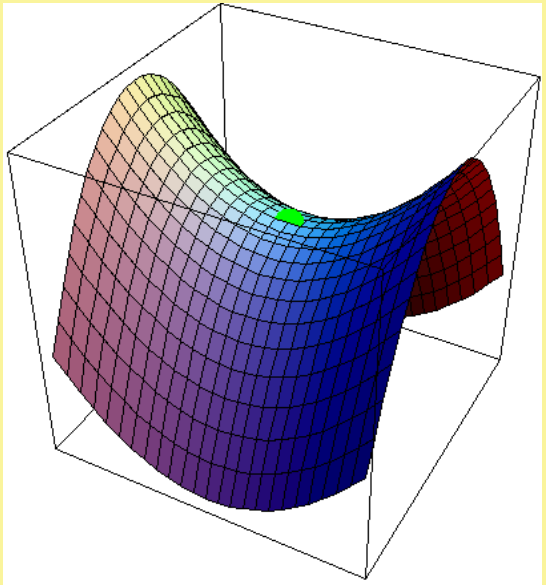


Gradient Descent Challenges

- 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 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.



Optimizing Gradient Descent

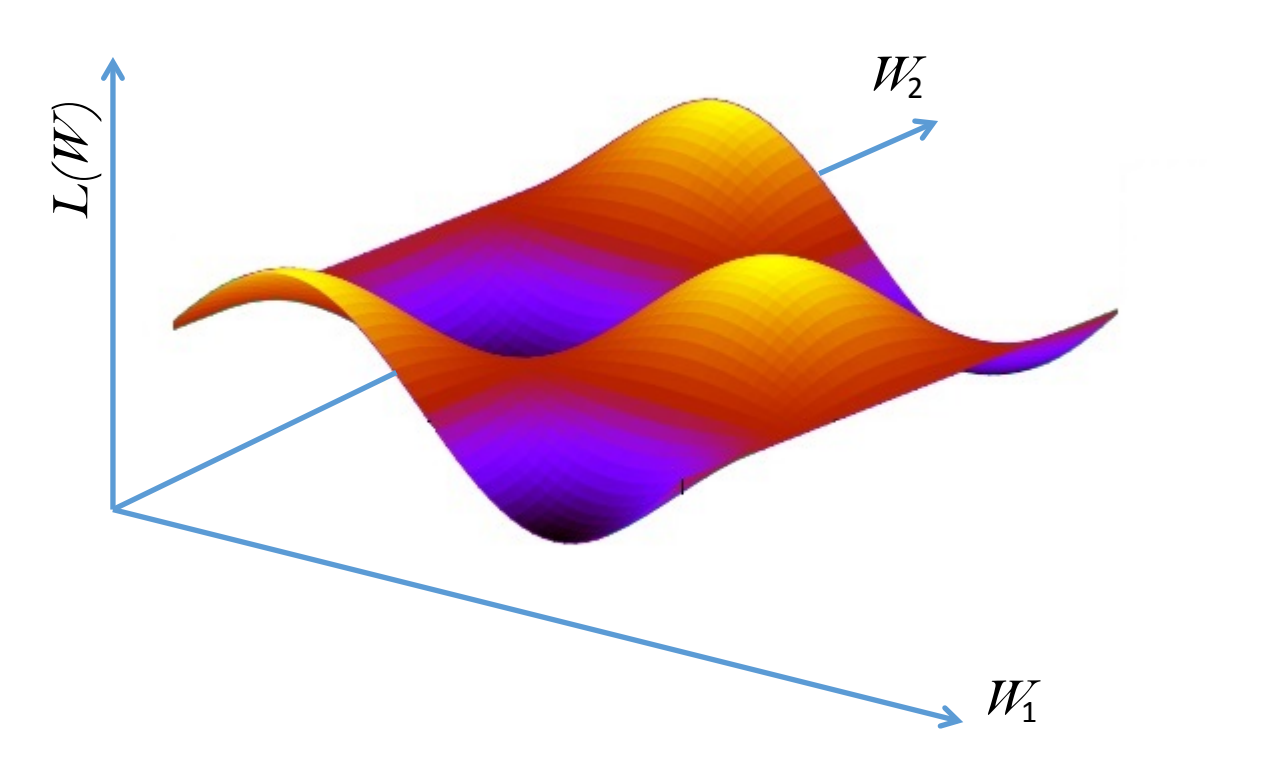
CONCEPTS COVERED

Concepts Covered:

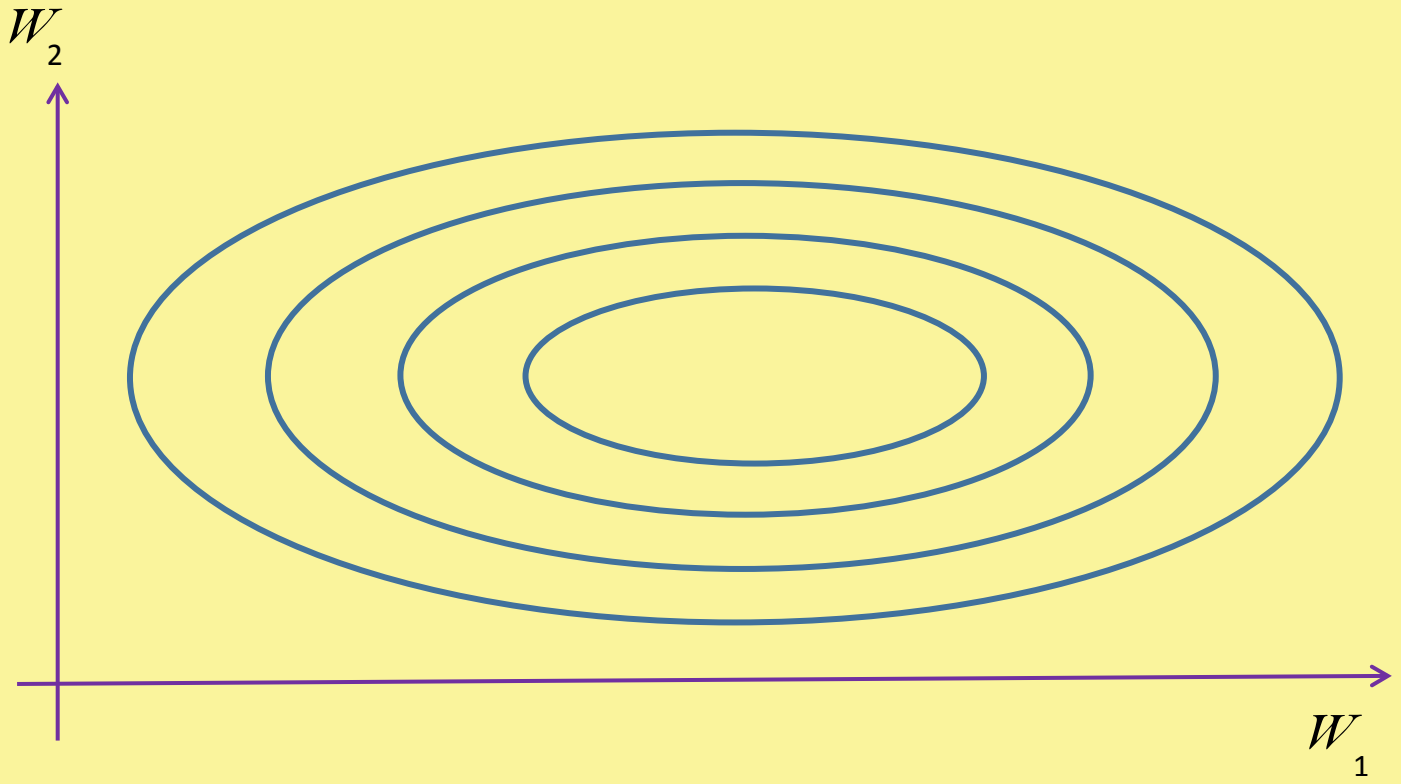
- CNN
  - ResNet
  - Gradient Descent Challenges
  - Momentum Optimizer
  - Adagrad.
  - etc.

Momentum Optimizer

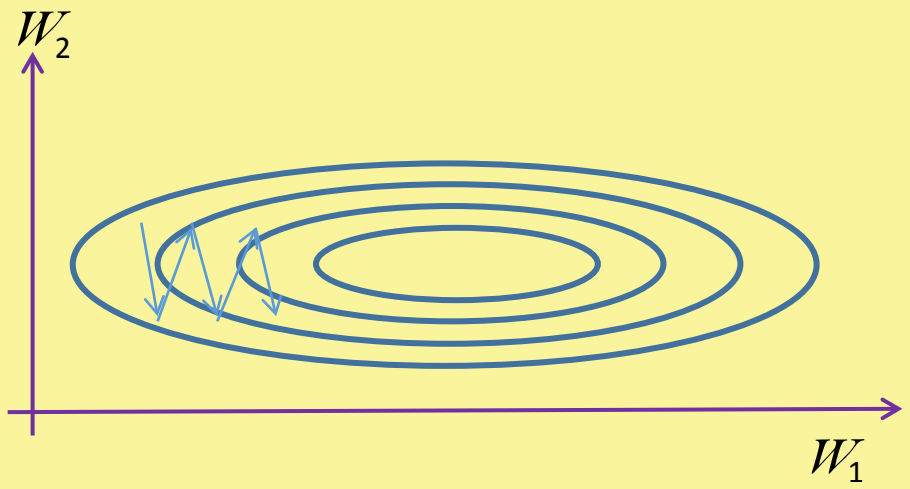
Momentum Optimizer



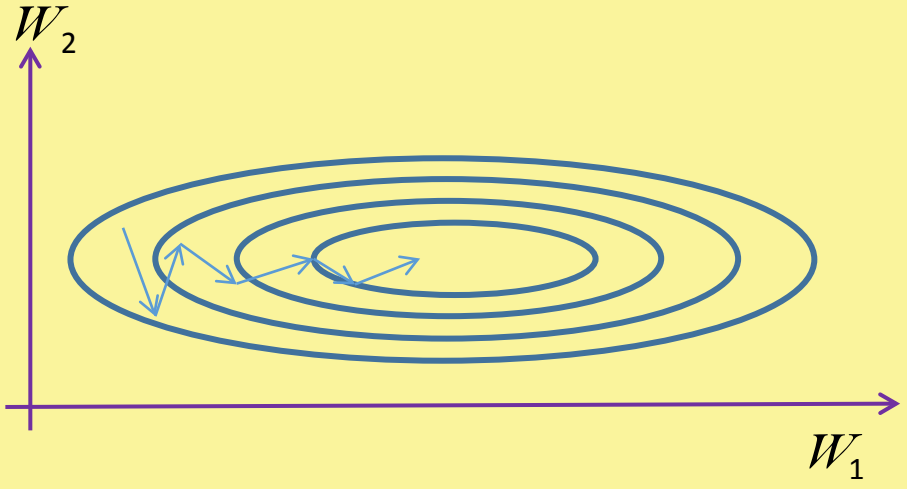
Momentum Optimizer



Momentum Optimizer



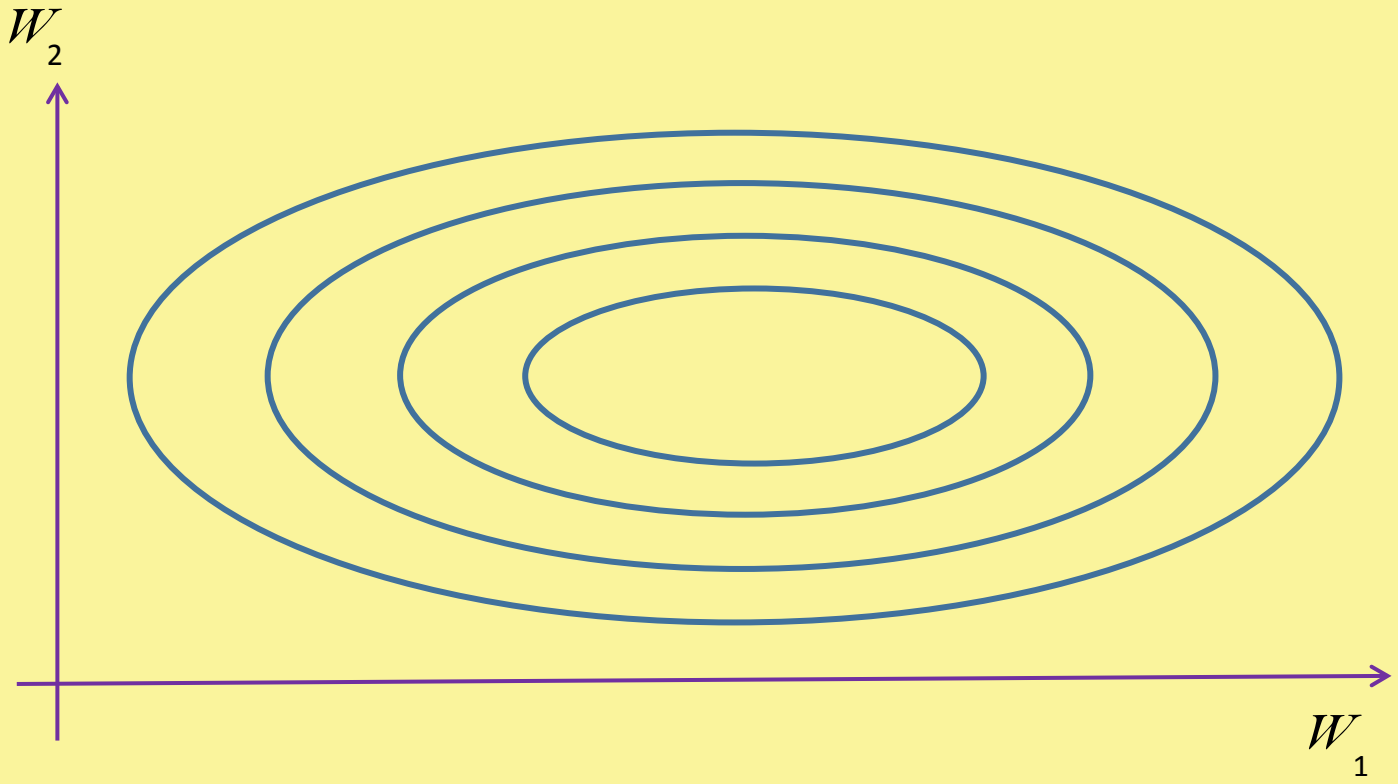
SGD



SGD with Momentum

Nesterov Accelerated Gradient (NAG)

Nesterov Accelerated Gradient (NAG)



Problem with Momentum Optimizer/NAG

- Both the algorithms require the hyper-parameters to be set manually.
- These hyper-parameters decide the learning rate.
- The algorithm uses same learning rate for all dimensions.
- The high dimensional (mostly) non-nonconvex nature of loss function may lead to different sensitivity on different dimension.
- We may require learning rate could be small in some dimension and large in another dimension.

NPTEL ONLINE CERTIFICATION COURSES

Thank you