Optimizing Neural Networks

Agenda

- Neural Networks Quiz
- Advanced Optimizers beyond SGD
- Dropout for Regularization
- Batch Normalization for Regularization
- Weight Initialization Techniques

Let's begin the discussion by answering a few questions on neural networks

Neural Networks Quiz

What does momentum in SGD with Momentum help achieve?

A. Accelerates convergence using past gradients

B. Adjusts learning rates adaptively

C. Reduces the variance of gradient updates

D. Prevents overshooting during optimization

Neural Networks Quiz

Great Learning
POWER AHEAD

What does momentum in SGD with Momentum help achieve?

A. Accelerates convergence using past gradients

B. Adjusts learning rates adaptively

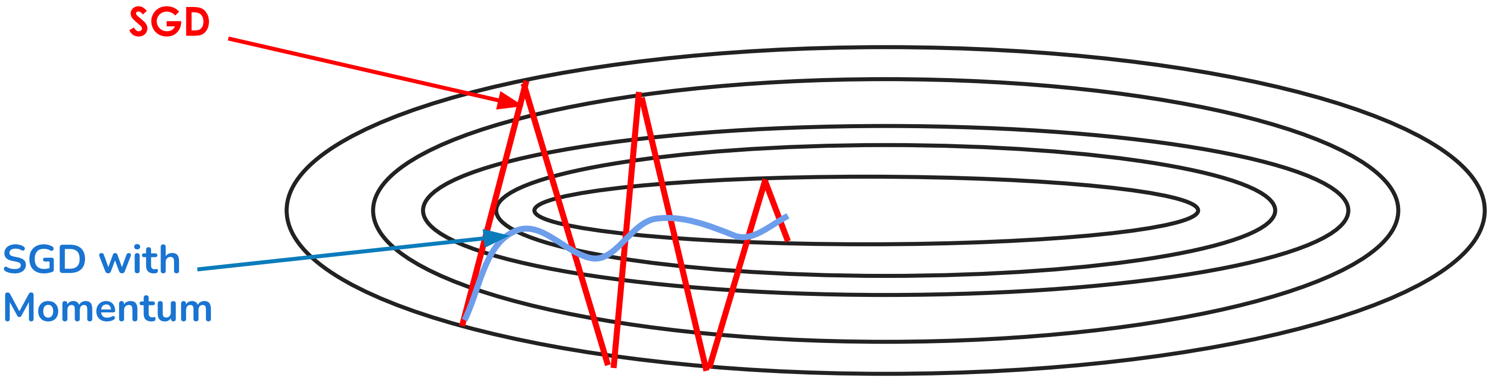
C. Reduces the variance of gradient updates

D. Prevents overshooting during optimization

Stochastic Gradient Descent with Momentum

SGD with momentum **accelerates convergence** by incorporating past gradients to maintain a consistent direction towards the minima

Momentum enhances this process by **reducing oscillations** and **facilitating smoother optimization trajectories**



Neural Networks Quiz

Which statement correctly describes how Adam and SGD with Momentum handle the learning rate?

A. The learning rate is constant throughout training for both Adam and SGD with Momentum.

B. Adam adjusts the learning rates separately for each parameter, while SGD with Momentum maintains a fixed learning rate.

C. SGD with Momentum changes the learning rate depending on gradient magnitude, whereas Adam sticks to a consistent learning rate.

D. Both Adam and SGD with Momentum utilize a dynamic learning rate approach throughout the training process.

Neural Networks Quiz

Which statement correctly describes how Adam and SGD with Momentum handle the learning rate?

A. The learning rate is constant throughout training for both Adam and SGD with Momentum.

B. Adam adjusts the learning rates separately for each parameter, while SGD with Momentum maintains a fixed learning rate.

C. SGD with Momentum changes the learning rate depending on gradient magnitude, whereas Adam sticks to a consistent learning rate.

D. Both Adam and SGD with Momentum utilize a dynamic learning rate approach throughout the training process.

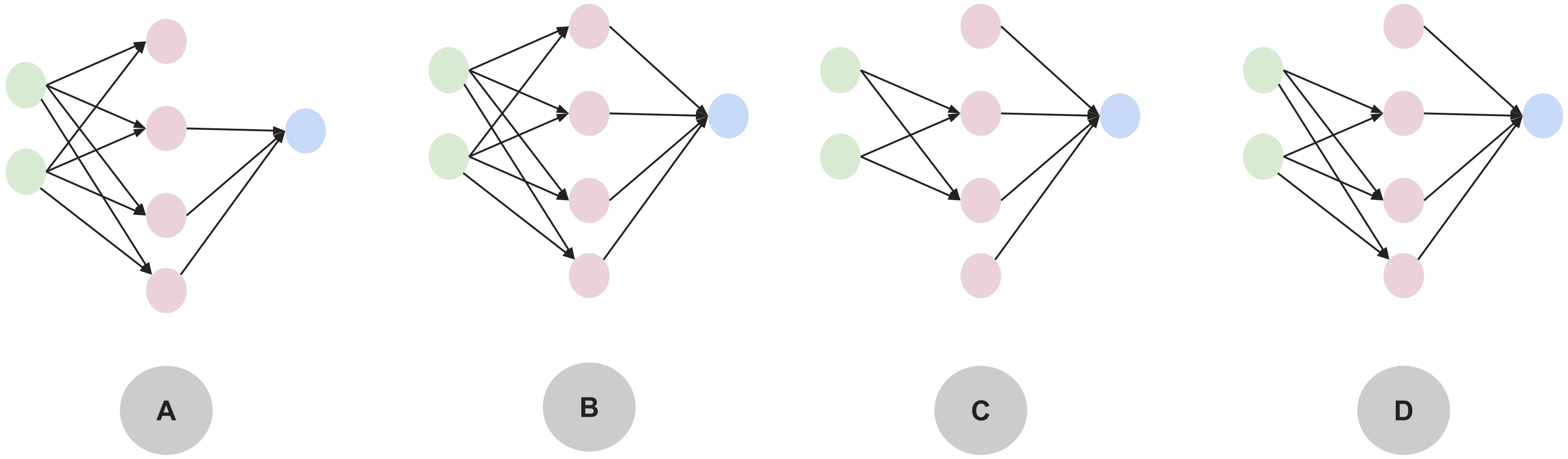
Learning Rates in ADAM & SGD with momentum

SGD with Momentum	Adam
Stochastic Gradient Descent with Momentum	Adaptive Moment Estimation
Adds a momentum component	Adds a momentum component
Fixed Learning Rate	Adaptive Learning Rate

The learning rate is different for each model parameter and depends on the value of the gradient

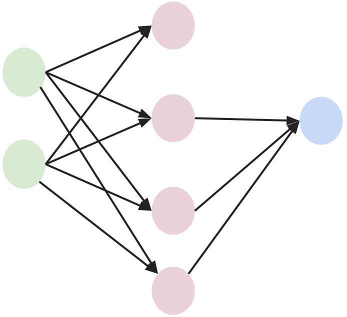
Neural Networks Quiz

Which of the following neural networks best represents a dropout rate of 0.5?

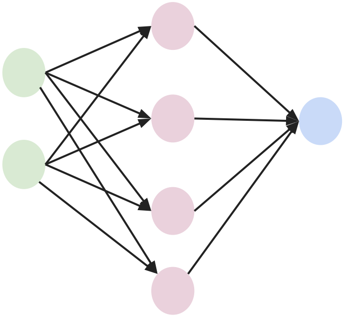


Neural Networks Quiz

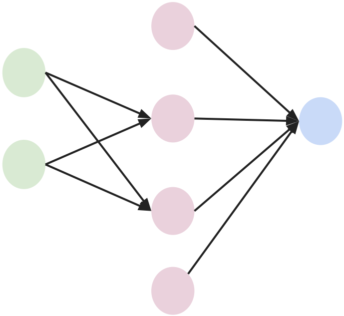
Which of the following neural networks best represents a dropout rate of 0.5?



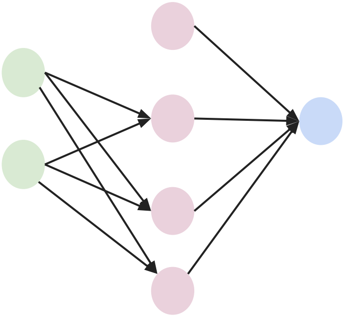
A



B



C



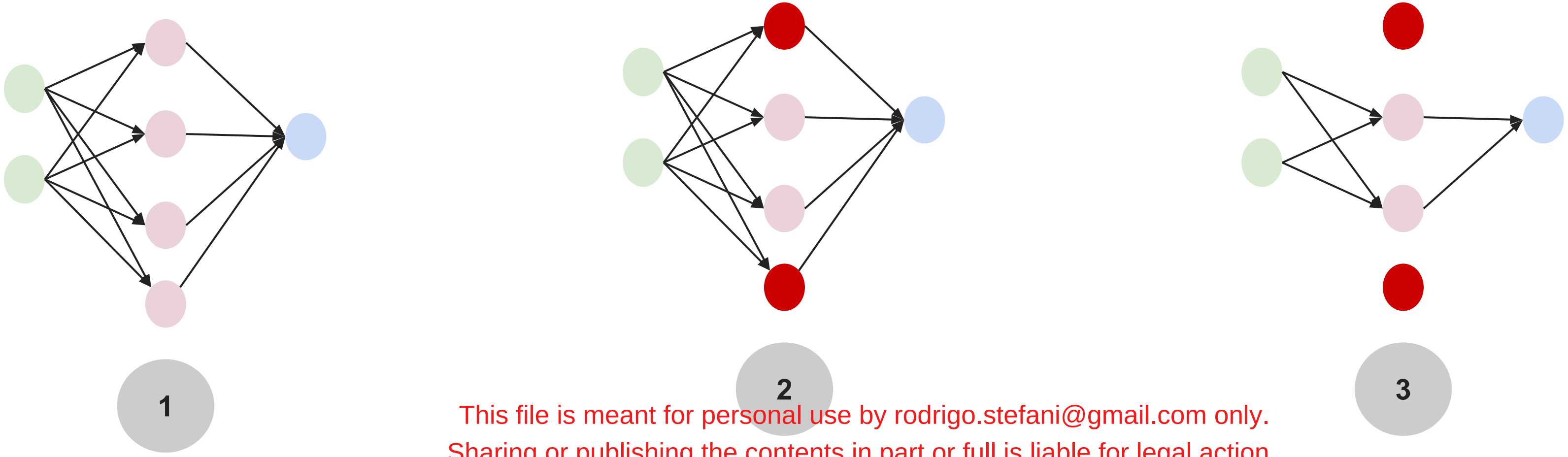
D

Dropout

Step 1 - Choose a dropout rate p (p = 0.5 here)

Step 2 - Randomly select 100*p% of the neurons (50% here)

Step 3 - Deactivate the selected neurons by setting them to zero.



Neural Networks Quiz

How does Dropout contribute to dealing with overfitting in neural networks?

A. Decreasing the complexity of the neural network

B. Increasing the complexity of the neural network

C. Changing the activation function of a neuron

D. By adding Gaussian noise to the input

Neural Networks Quiz

How does Dropout contribute to dealing with overfitting in neural networks?

A. Decreasing the complexity of the neural network

B. Increasing the complexity of the neural network

C. Changing the activation function of a neuron

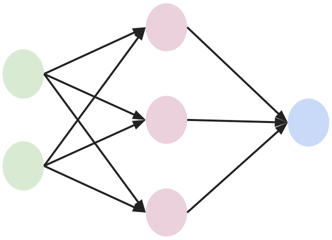
D. By adding Gaussian noise to the input

Dropout for Regularization

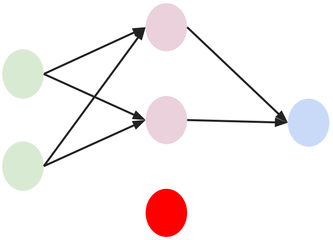
Dropout prevents overfitting by randomly deactivating neurons during training

Independence: Fosters independent learning among neurons

Ensemble: Creates a diverse ensemble of networks as neurons are deactivated randomly during training and the final result is an average prediction



Before Dropout



After Dropout

Neural Networks Quiz

How many learnable parameters are there in a batch normalization layer?

A. 4

B. 3

C. 0

D. 2

Neural Networks Quiz

How many learnable parameters are there in a batch normalization layer?

- A 4
- B 3
- C 0
- D 2

Batch Normalization - Working

Step 1 - Normalization: Normalize X (input) by subtracting its mean and dividing by its standard deviation.

No learnable parameters in Step 1

Step 2 - Scaling: Scale the output of Step 1 by multiplying it with a learnable parameter *gamma*

Step 3 - Shifting: Shift the output of Step 2 by adding an offset (learnable parameter *beta*)

Neural Networks Quiz

Which of the following accurately describes the purpose of batch normalization in neural networks?

A. Minimizing overfitting by adding noise to the input data.

B. Preventing vanishing gradients by initializing weights appropriately.

C. Reducing internal covariate shift by normalizing layer activations.

D. Decreasing model complexity by adding more parameters.

Neural Networks Quiz

Which of the following accurately describes the purpose of batch normalization in neural networks?

A. Minimizing overfitting by adding noise to the input data.

B. Preventing vanishing gradients by initializing weights appropriately.

C. Reducing internal covariate shift by normalizing layer activations.

D. Decreasing model complexity by adding more parameters.

Batch Normalization - Purpose

Stabilizing Training: Reduces internal covariate shift.

Regularization: Acts as a form of regularization.

Improved Gradient Flow: Enhances gradient flow for faster convergence.

Neural Networks Quiz

How does weight initialization contribute to improved learning in a neural network?

A. By introducing uniform gradients.

B. By ensuring symmetric neuron behavior.

C. By adjusting initial weights.

D. By hindering adaptation.

Neural Networks Quiz

How does weight initialization contribute to improved learning in a neural network?

A. By introducing uniform gradients.

B. By ensuring symmetric neuron behavior.

C. By adjusting initial weights.

D. By hindering adaptation.

Weight Initialization

Sets up the starting point for learning in a neural network

Helps break symmetry among neurons

Prevents gradient problems

Makes learning faster and more efficient overall

Neural Networks Quiz

What is the common problem that can arise when weights are initialized randomly in a neural network?

A. Vanishing Gradients

B. Exploding Gradients

C. Vanishing or Exploding Gradients

D. Learning rate becoming too high

Neural Networks Quiz

What is the common problem that can arise when weights are initialized randomly in a neural network?

A. Vanishing Gradients

B. Exploding Gradients

C. Vanishing or Exploding Gradients

D. Learning rate becoming too high

Weight Initialization

Randomly initializing weights in a neural network can lead to **two critical problems during training**

Vanishing gradients: Occurs when the gradients (derivatives) become extremely small, slowing down or halting the learning process

Exploding gradients: Occurs when the gradients become excessively large, causing instability and hindering effective learning

Specific weight initialization strategies help overcome these problems

Xavier (or Glorot) Initialization - sigmoid and tanh activations

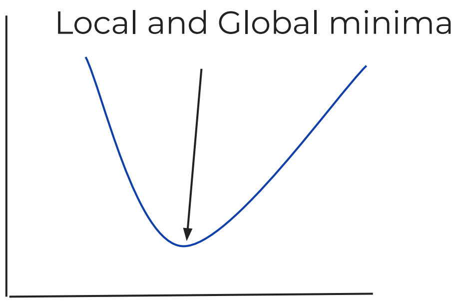
He Initialization - ReLU and variants of ReLU

Happy Learning !

APPENDIX

Need for Advanced Optimizers

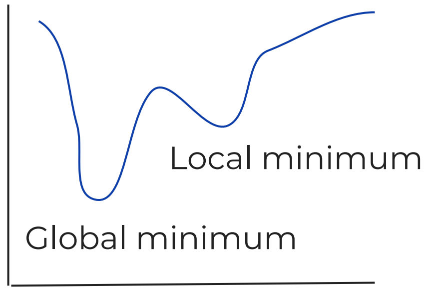
Convex Optimization



Efficient optimization, Global convergence

Guaranteed global convergence under conditions

Non-Convex Optimization



Complex, Local optima, Challenging optimization

No guarantee, Prone to local optima