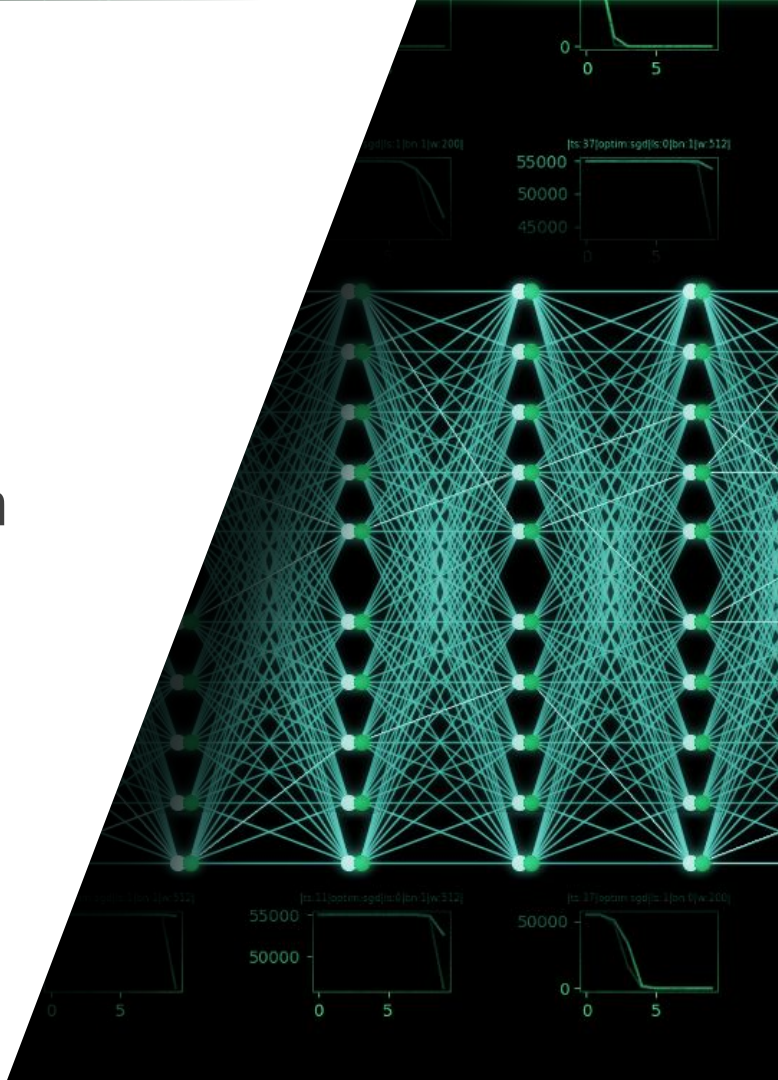
Pre-interpolation Loss Behavior in Neural Networks

Venter, Theunissen & Davel

Willem Venter

North-West University, South Africa
Centre for Artificial Intelligence Research

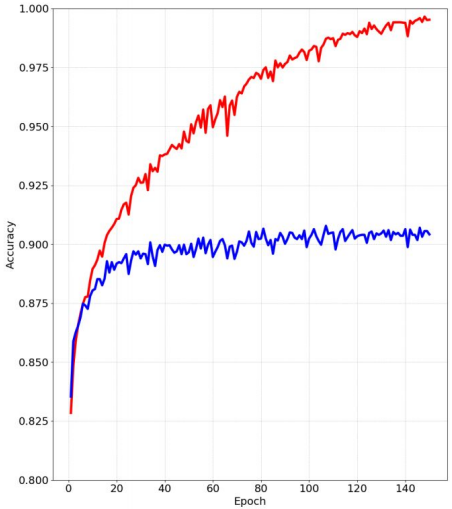


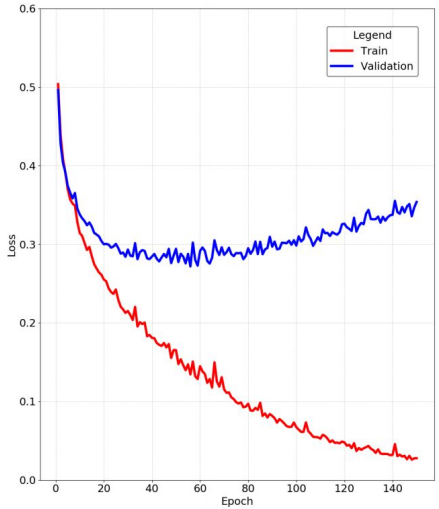
Overview

- Paradoxical relationship between validation loss and accuracy:
  - Limits of a point estimator
  - Overfitting on selected validation samples
- Empirical evidence
- Implications on generalization

Introduction

- Witnessed phenomenon:
  - Increase in both validation loss and accuracy
- Shallow local optima and borderline cases





- Evaluating using individual validation loss distributions:
  - Validation samples with large loss contributions

Approach

- Fully-connected feedforward networks using MLP architecture
- Omitted any regularization
- Trained till convergence

- First step
  - Determine whether phenomenon occurs in general scenarios
- Second step
  - Select best models for further analysis

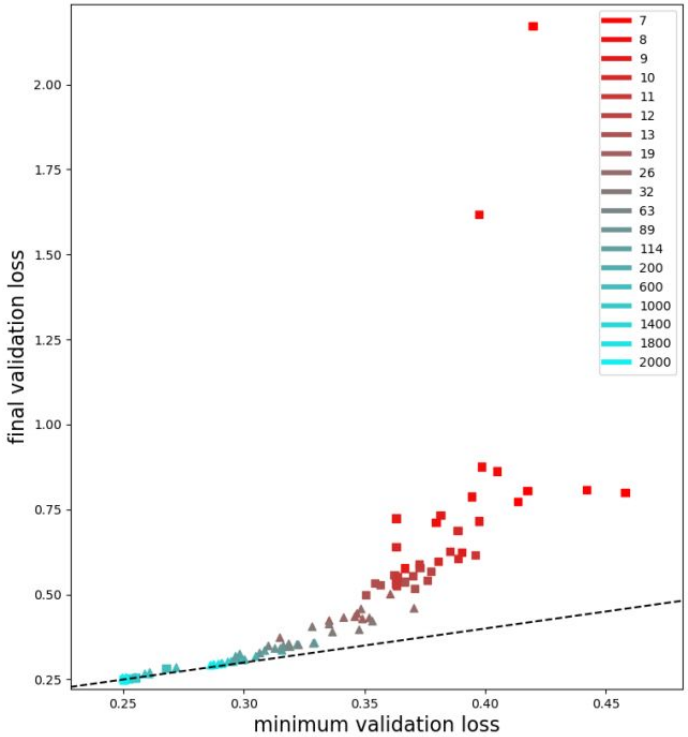
Architecture

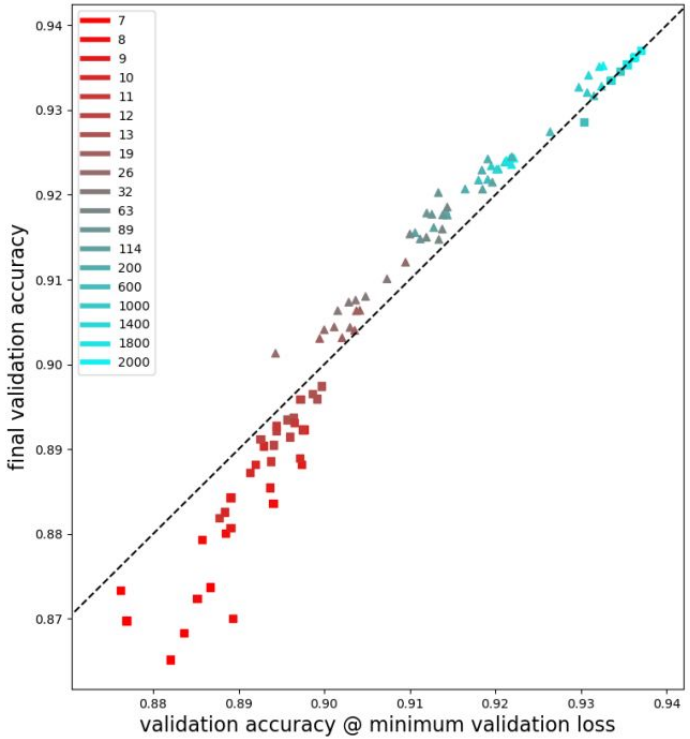
- Cross-entropy loss function
- Equal number of nodes per hidden layers
- Varied hyperparameters values:
  - Training and validation set sizes
  - Number of hidden layers
  - Number of nodes
  - Mini-batch sizes
  - Datasets (MNIST or FMNIST)
  - Optimizers (Adam or SGD)

Experimental Setups

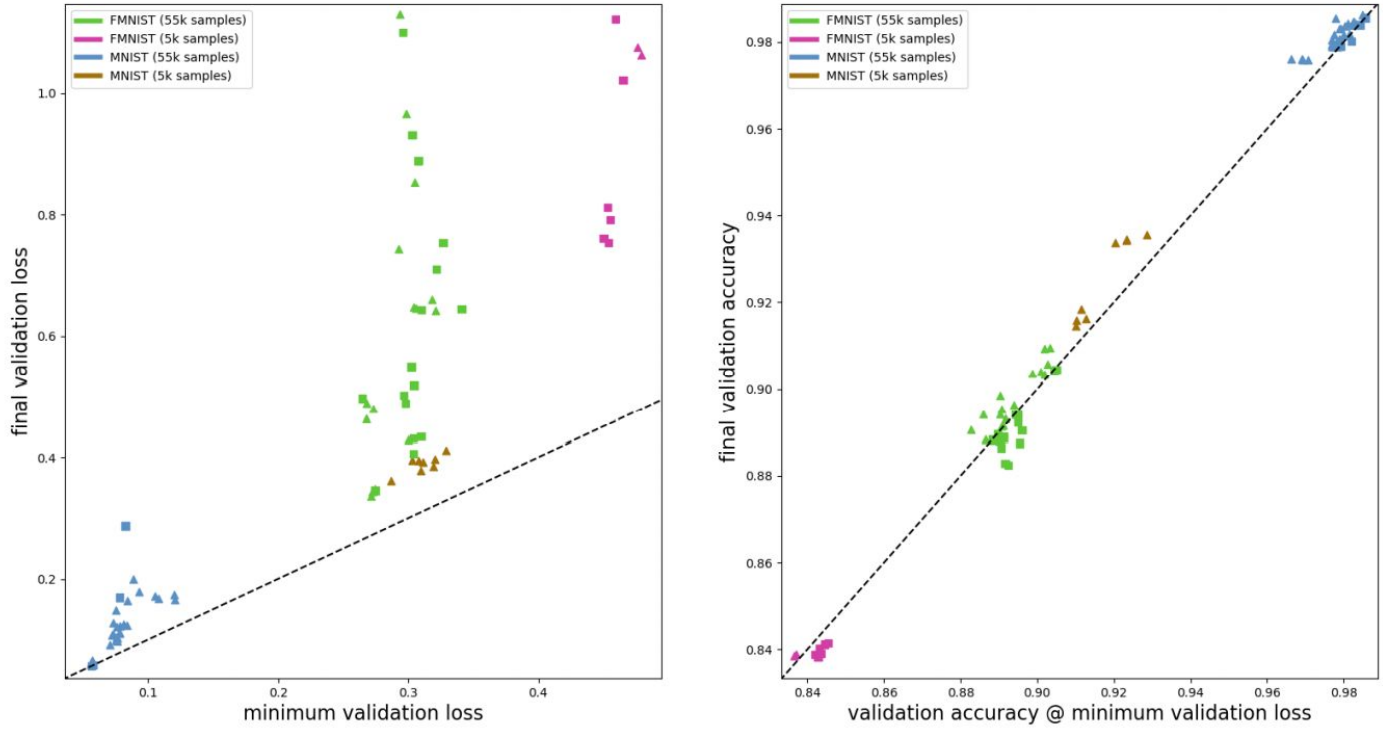
- First Experiment (MNIST):
  - 95 two-layer MLPs with 7 – 2000 nodes per layer
  - 5000 training samples
  - Mini-batch size of 64
  - Adam (57) and SGD (38) optimizers
- Second Experiment (MNIST, FMNIST):
  - 80 MLPs with varying depth (1 /3 /10 layers) and width (100/ 1000 nodes)
  - 5000 /55000 training samples
  - Mini batch sizes of 16/ 64/ 256
  - Adam and SGD optimizers

Results – First Experiment

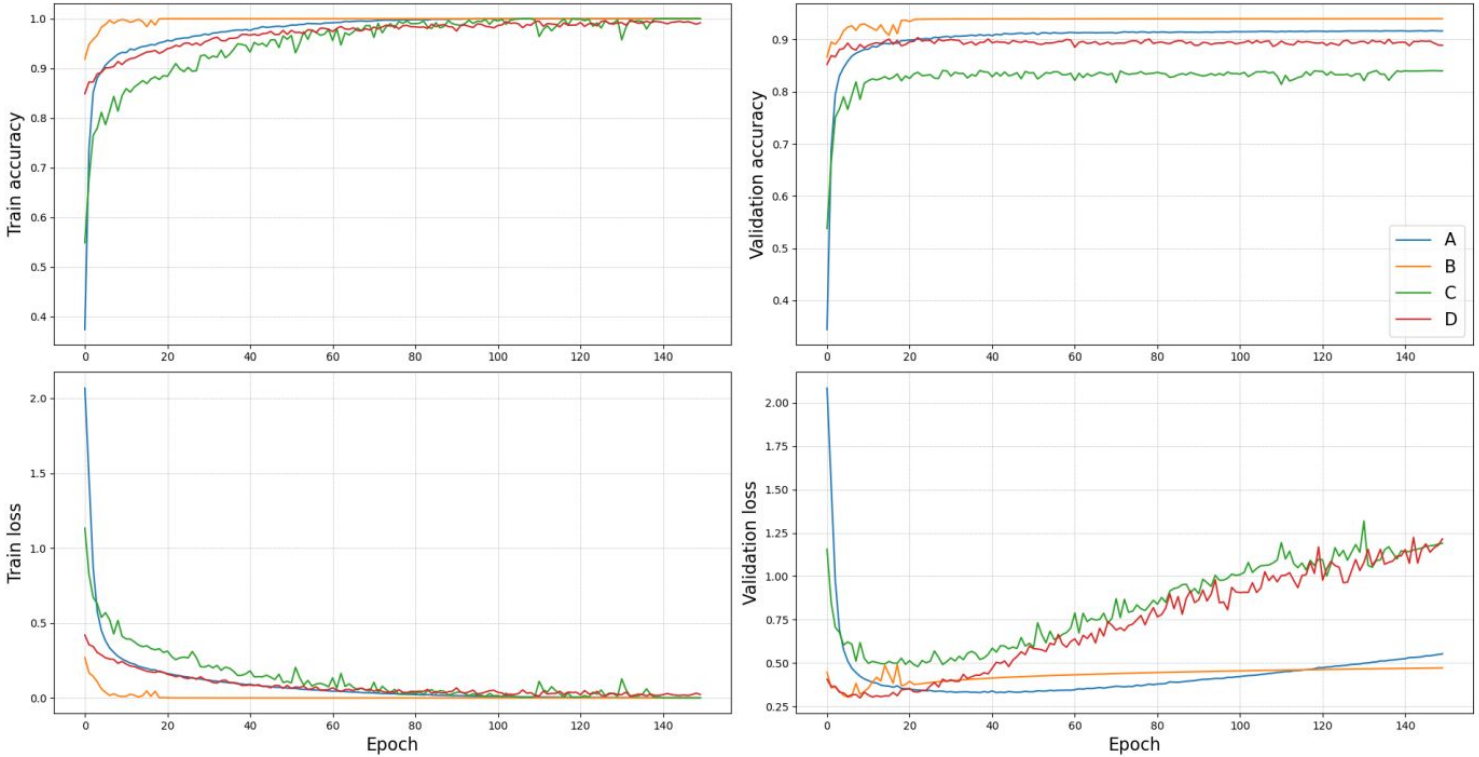




Results – Second Experiment



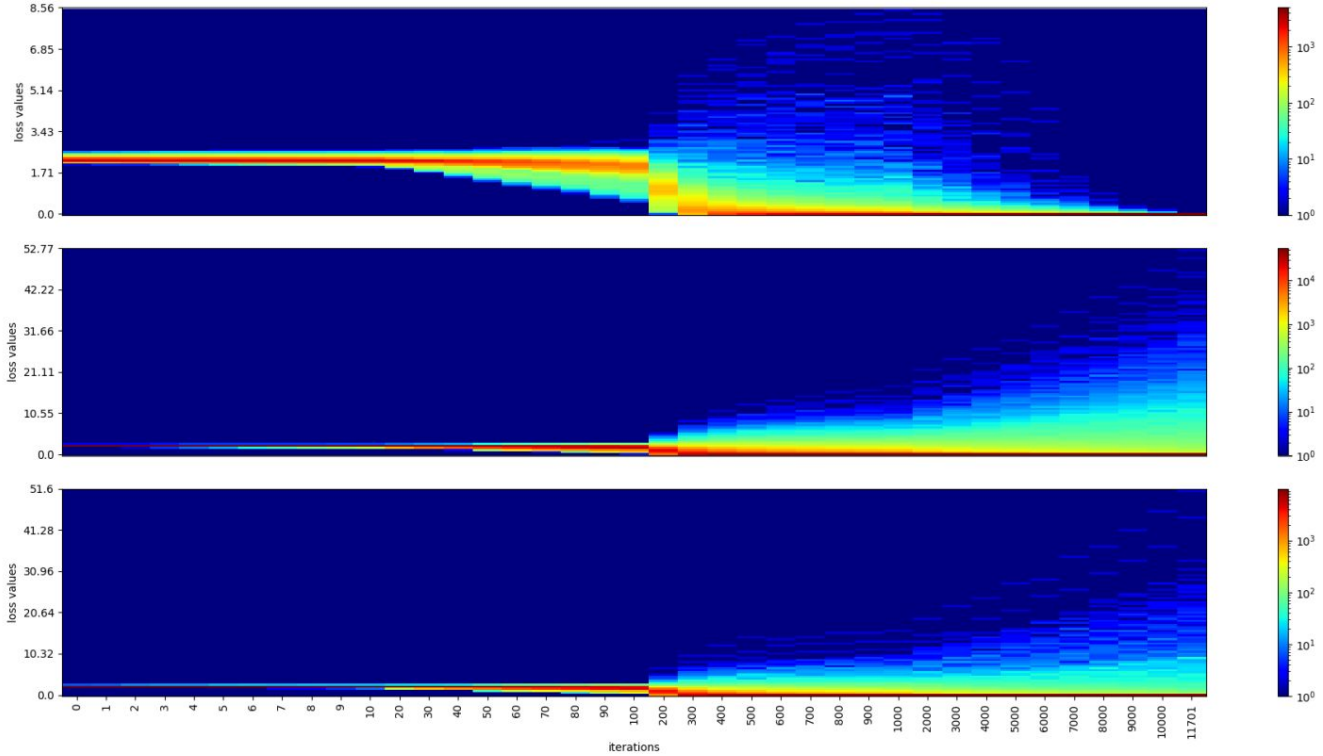
Results – Distribution analysis



Results – Distribution analysis

Model A:

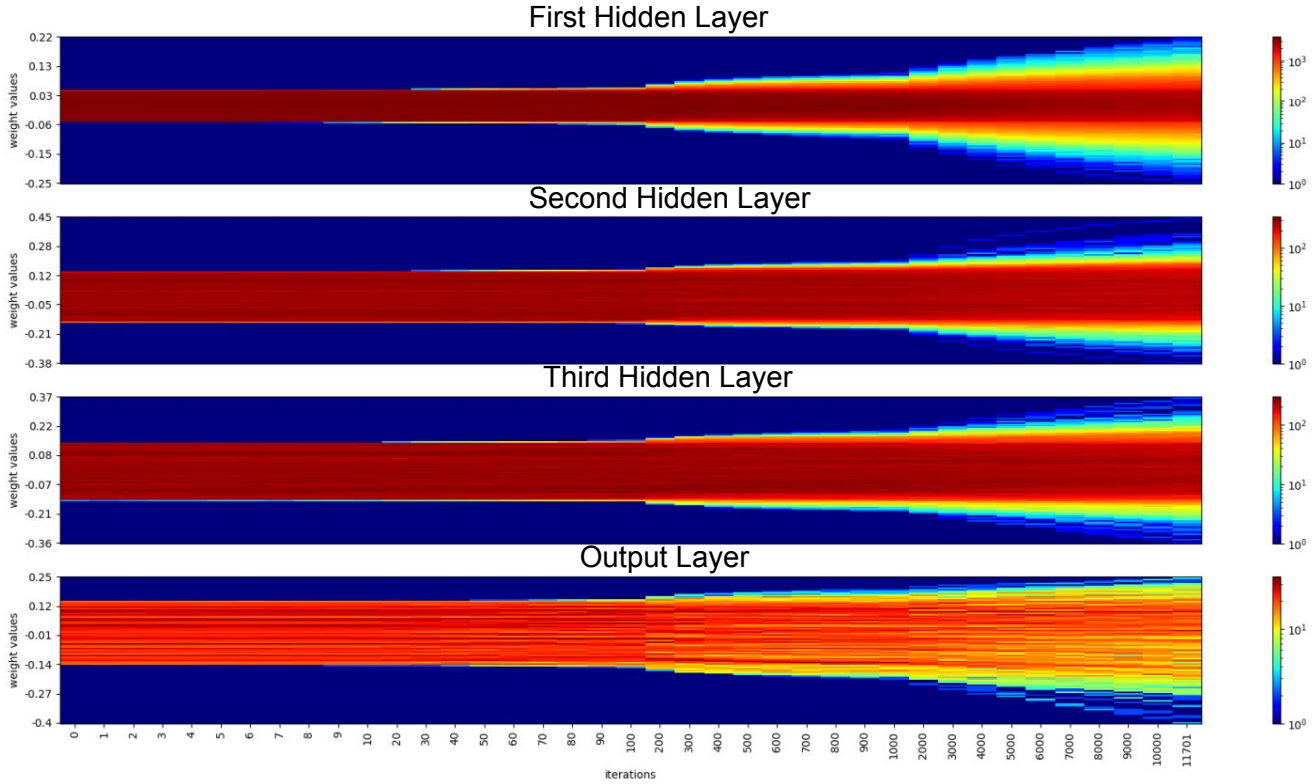
- 3x100 MLP
- 5000 MNIST samples
- Adam optimizer
- Mini-batch of 64



Results – Distribution analysis

Model A:

- 3x100 MLP
- 5000 MNIST samples
- Adam optimizer
- Mini-batch of 64

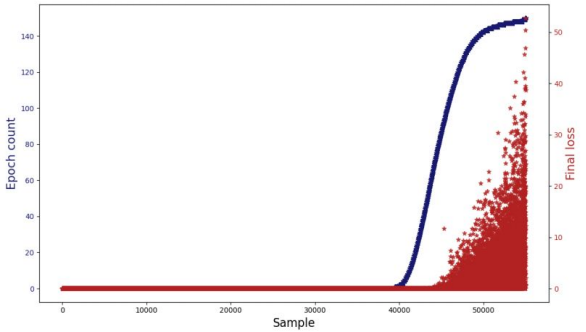


Results - Outliers

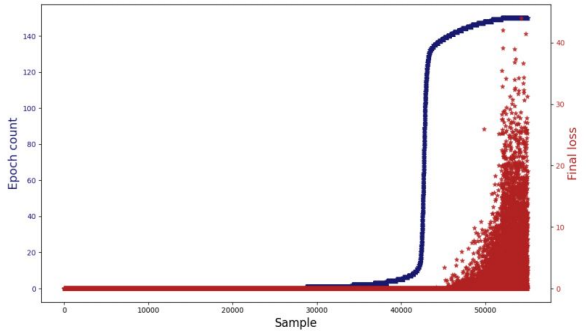
$Q_3 + 1.5 \times (Q_3 - Q_1)$

Q_3 - Third quartile of loss values

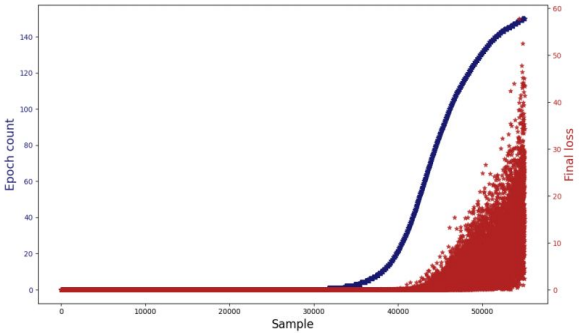
Q_1 - First Quartile of loss values

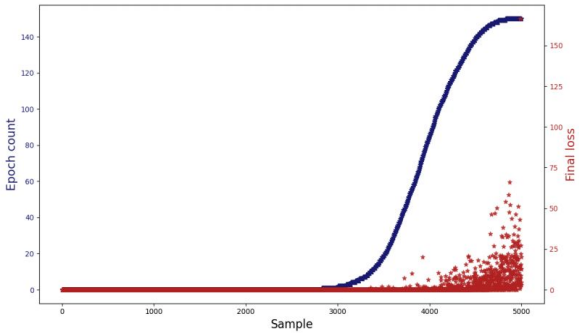


(a) A: Fitting 5k MNIST; Adam



(b) B: Fitting 5k MNIST; SGD





In summary

- Discrepancies between validation loss and accuracy can exist where both will show increases their respective metrics
- This is due to the large loss contributions of a minority of validation loss samples
- Point estimators are a poor estimate of generality because loss values aren’t necessarily close to a mean value
- Weight values are increased to minimize training loss but sacrifices generality in the process
- DNNs seem to be distributed and heterogenous and should be considered when focussing on generalization

Thank you!