CS550 Homework 1

Berat Biçer

January 2, 2020

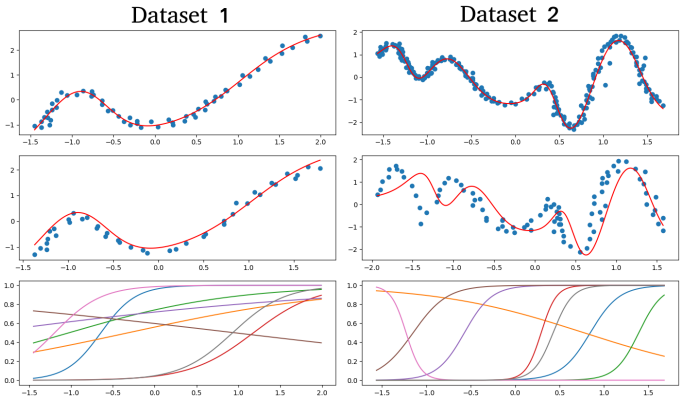
Network Configurations

Table 1: Network configuration for ANN1 and ANN2

Component	ANN1	ANN2
Hidden units used	8	8
Activation Function	Sigmoid	Sigmoid
Loss Function	SSE	SSE
Learning Rate	$5 * 10^{-3}$	$5 * 10^{-3}$
Initial Weights	$N(\mu = 0, \Sigma = 1)$	$N(\mu = 0, \Sigma = 1)$
# Epochs	49100	84200
Stopping Criteria (T, E)	$(10^{-8}, 50000)$	$(10^{-8}, 100000)$
Momentum	No	No
Normalization	Yes	Yes
Stochastic or Batch	Stochastic	Stochastic
Training Loss	0.010	0.025
Test Loss	0.031	0.379

The configurations of the networks are shared in Table 1. These configurations are not optimized, but the networks yield satisfactory results in test data, as shown in Figure 1.

Figure 1: Network plots for Section 1. Top row shows the network on train sets and middle row on test sets where last row shows the values of hidden units.



Experiments

Part C

Table 2: Configuration for Dataset 1 in Part C.

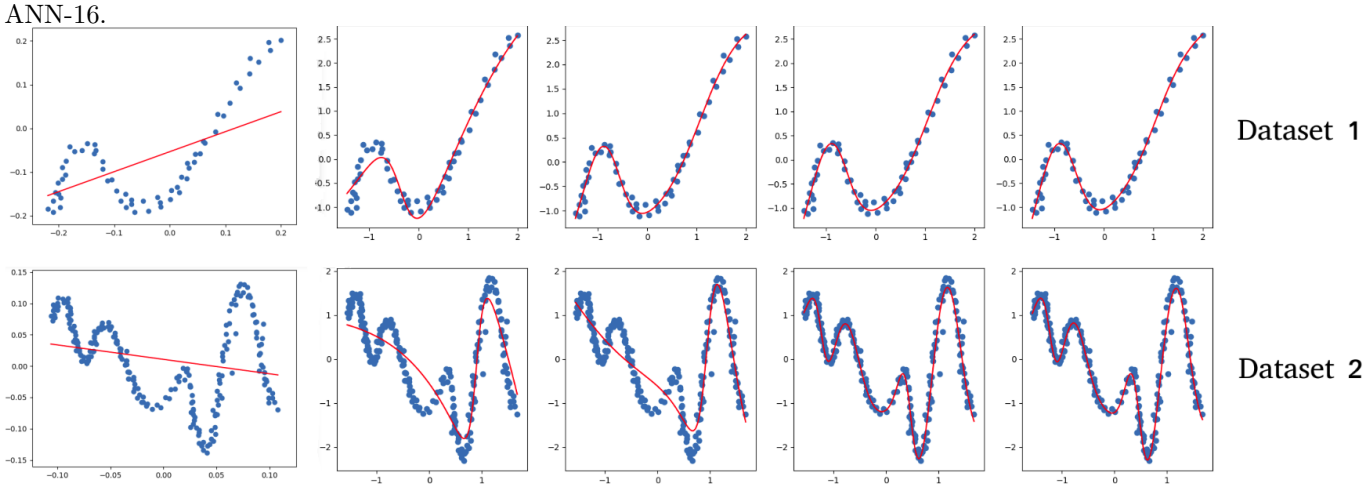
Component	Lin.Reg.	ANN2	ANN4	ANN8	ANN16
Hidden Units	0	2	4	8	16
Activation	NA	Sigmoid			
Loss Function	SSE				
Learning Rate	0.008	0.005			
# Epochs	180	50k	50k	46.5k	50k
Initial Weights	$N(\mu = 0, \Sigma = 1)$				
Stopping Criteria	$(10^{-8}, 180)$	$(10^{-8}, 50k)$			
Momentum	No				
Normalization	Yes				
Stochastic or Batch	Batch	Stochastic			
Train Loss	0.185	0.0105	0.0104	0.0107	0.0109
Test Loss	0.24	0.0352	0.0311	0.0307	0.0309

Network configurations are given in Table 2 & 3, and the network plots are given in Figure 2. This experiment focuses on the effect of change in number of hidden units on network performance. Note that initial weights are randomly selected at each time, which theoretically affect the learnt parameters. Starting with dataset 1, linear regression performs worse by a large margin due to the nonlinear nature of the data and the models inability to cope with it. Comparing test losses, since the learned approximations are similar, the results are similar except ANN-2. This is likely because two sigmoids aren't enough to capture the data trend. We also observe the number of train epochs stay approximately the same, meaning the learning rate and possibly momentum values can be further optimized for a better stopping criteria. For ANN16, however, we see test loss increases slightly. While it is possible that this increase is coincidental, I suspect the network overfits due to high degree of freedom caused by 16 sigmoids.

Table 3: Configuration for Dataset 2 in Part C.

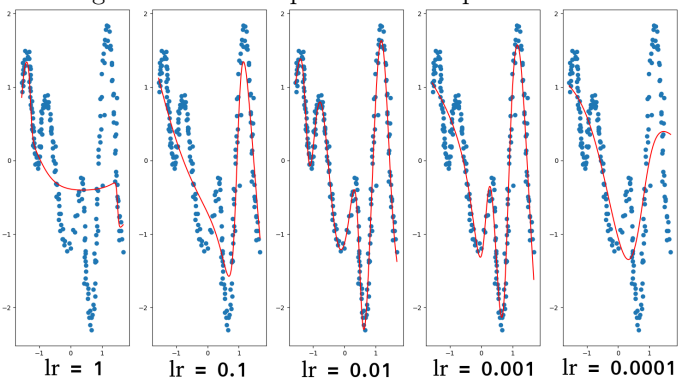
Component	Lin.Reg.	ANN2	ANN4	ANN8	ANN16
Hidden Units	0	2	4	8	16
Activation	NA	Sigmoid			
Loss Function	SSE				
Learning Rate	0.008	0.0075			
# Epochs	180	40.1k	34.5k	88.1k	100k
Initial Weights	$N(\mu = 0, \Sigma = 1)$				
Stopping Criteria	$(10^{-8}, 180)$	$(10^{-8}, 100k)$			
Momentum	No				
Normalization	Yes				
Stochastic or Batch	Batch	Stochastic			
Train Loss	0.438	0.109	0.085	0.0256	0.0262
Test Loss	0.49	0.3364	0.3229	0.3772	0.4014

Figure 2: Network plots for for experiment C. From left to right: plots for linear regression, ANN-2, ANN-4, ANN-8, ANN-16.



Studying the hidden unit plots, we observed that number of sigmoids converging to zero (and having no contribution, ie. saturating) increases as the number of hidden units increase. For ANN-2, 4, 8, 16 these numbers are 1, 3, and 9. These discussions are also valid for dataset 2. Differences are due to 7 local maximas, so ANN-2 and ANN-4 performs significantly worse than ANN-8 and ANN-16 due to insufficient model capacity.

Figure 3: Network plots for for experiment D.



Part D

Network configurations are shared in Table 4, and the resulting networks are plotted in Figure 3. The stopping criteria is selected so that the networks overfit. From the plots, we see learning rate 0.01 learns the data trend well. For 1 and 0.1, the optimization oversteps the optimal point and the learned networks fail to capture the data trend. For 0.001 and 0.0001, we suspect that if the stopping criteria is relaxed in terms of maximum number of epochs, learned networks would be similar.

Table 4: Configurations for Part D.

Learning Rate	1	0.1	0.01	0.001	0.0001
Hidden Units	8				
Activation	Sigmoid				
Loss Function	SSE				
# Epochs	100k	100k	100k	100k	100k
Initial Weights	$\mathrm{N}(\mu = 0, \Sigma = 1)$				
Stopping Criteria	$(10^{-8}, 100k)$				
Momentum	No				
Normalization	Yes				
Stochastic or Batch	Stochastic				
Train Loss	0.3544	0.102	0.0261	0.0561	0.1061
Test Loss	0.4635	0.3418	0.3212	0.320	0.382

Part E

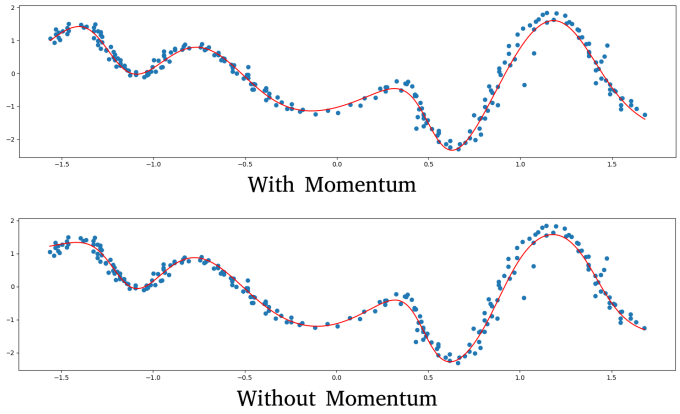
Table 5: Configurations for Part E.

Component	ANN-NM	ANN-M
Hidden Units	8	
Activation	Sigmoid	
Loss Function	SSE	
Epochs	88.1k	74k
Learning Rate	0.0075	
Initial Weights	$\mathrm{N}(\mu = 0, \Sigma = 1)$	
Stopping Criteria	$(10^{-8}, 100k)$	
Momentum	0	0.9
Normalization	Yes	
Stochastic or Batch	Stochastic	
Train Loss	0.0256	0.0243
Test Loss	0.3772	0.4172

Network configurations and results of the experiment can be found in Table 5, and the network plots are in Figure 4. We can see that the network with momentum achieves comparative results with less epochs. The value for the

momentum factor is also important. Here, momentum value of 0.9. We see that network with higher momentum converges faster but the obtained test loss increases, meaning that the learning is not stable.

Figure 4: Network plots for for experiment E.



Part F

Network configurations are given in Table 6, and the network plots in Figure 5. Batch learning takes longer to train than stochastic gradient descent. It it likely that if we increase the learning rate for batch learning, training may become faster. This implies that batch learning is not as stable as stochastic gradient descent.

Table 6: Configurations for Part F.

Component	ANN-Stochastic	ANN-Batch
Hidden Units	8	
Activation	Sigmoid	
Loss Function	SSE	
Epochs	80k	210k
Learning Rate	0.0075	
Initial Weights	$N(\mu = 0, \Sigma = 1)$	
Stopping Criteria	$(10^{-8}, 100k)$	$(10^{-8}, 250k$
Momentum	No	
Normalization	Yes	
Stochastic or Batch	Stochastic	Batch
Train Loss	0.0255	0.0253
Test Loss	0.3923	0.3724

Figure 5: Network plots for for experiment F.