Debug proof that my computed gradient matches the estimated gradient:

dWL\_cum =

Columns 1 through 11

0.0412 0.0606 0.0261 0.0311 0.0377 0.0247 0.0445 0.0458 0.0400 0.0537 0.0485

-0.0797 -0.1150 -0.0492 -0.0578 -0.0706 -0.0467 -0.0826 -0.0871 -0.0746 -0.1021 -0.0898

Columns 12 through 20

0.0478 0.0318 0.0398 0.0528 0.0269 0.0249 0.0285 0.0422 0.0584

-0.0900 -0.0606 -0.0751 -0.1003 -0.0513 -0.0462 -0.0534 -0.0803 -0.1103

est\_dWkj =

Columns 1 through 11

0.0412 0.0606 0.0261 0.0311 0.0377 0.0247 0.0445 0.0458 0.0400 0.0537 0.0485

-0.0797 -0.1150 -0.0492 -0.0578 -0.0706 -0.0467 -0.0826 -0.0871 -0.0746 -0.1021 -0.0898

Columns 12 through 20

0.0478 0.0318 0.0398 0.0528 0.0269 0.0249 0.0285 0.0422 0.0584

-0.0900 -0.0606 -0.0751 -0.1003 -0.0513 -0.0462 -0.0534 -0.0803 -0.1103

delta\_L\_cum =

0.0795

-0.1504

delta\_L\_est =

0.0795

-0.1504

dW\_Lminus1\_cum =

0.0034 0.0054

-0.0039 -0.0062

0.0003 0.0010

0.0037 0.0052

0.0006 0.0015

0.0066 0.0106

0.0030 0.0047

0.0053 0.0086

0.0010 0.0014

-0.0005 -0.0003

0.0008 0.0007

-0.0038 -0.0064

0.0043 0.0066

-0.0015 -0.0028

0.0012 0.0024

0.0029 0.0037

0.0018 0.0032

-0.0039 -0.0060

-0.0011 -0.0019

-0.0028 -0.0040

est\_dWji =

0.0034 0.0054

-0.0039 -0.0062

0.0003 0.0010

0.0037 0.0052

0.0006 0.0015

0.0066 0.0106

0.0030 0.0047

0.0053 0.0086

0.0010 0.0014

-0.0005 -0.0003

0.0008 0.0007

-0.0038 -0.0064

0.0043 0.0066

-0.0015 -0.0028

0.0012 0.0024

0.0029 0.0037

0.0018 0.0032

-0.0039 -0.0060

-0.0011 -0.0019

-0.0028 -0.0040

delta\_Lminus1\_cum =

0.0223

-0.0251

0.0051

0.0199

0.0069

0.0434

0.0192

0.0352

0.0056

-0.0004

0.0019

-0.0266

0.0265

-0.0124

0.0108

0.0135

0.0135

-0.0244

-0.0080

-0.0155

delta\_Lminus1\_est =

0.0223

-0.0251

0.0051

0.0199

0.0069

0.0434

0.0192

0.0352

0.0056

-0.0004

0.0019

-0.0266

0.0265

-0.0124

0.0108

0.0135

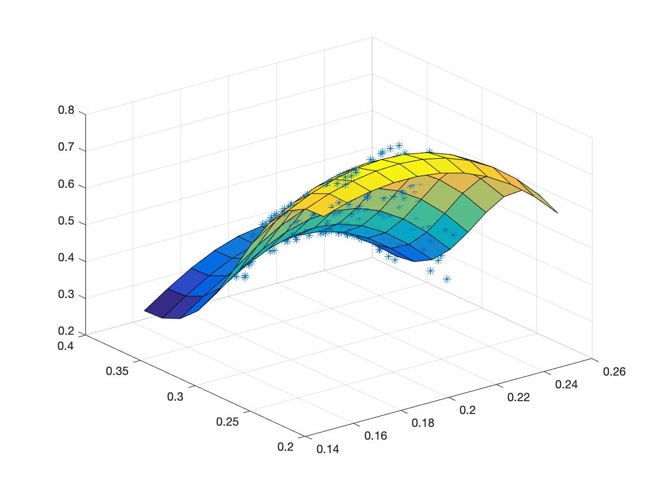
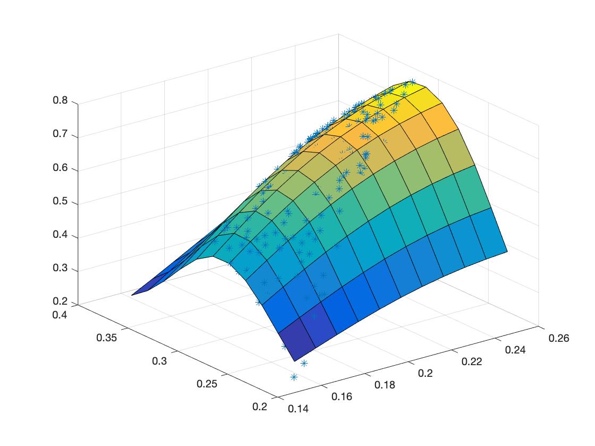
0.0135

-0.0244

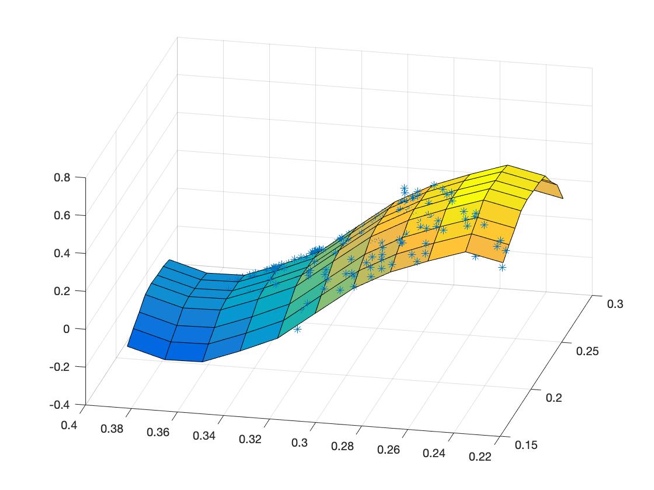
-0.0080

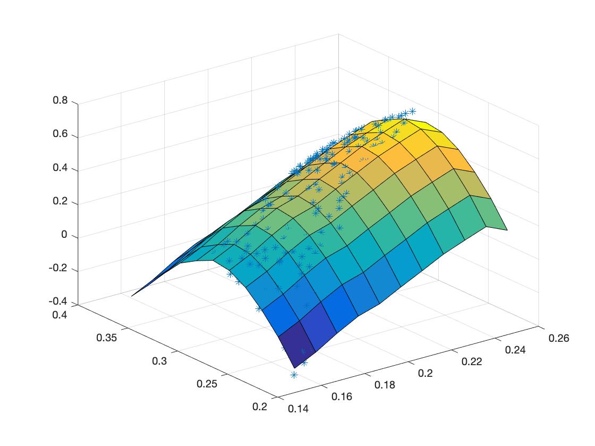
-0.0155

I first ran the non-preset network. As expected: convergence took a really long time. So long in fact that I forgot it was running. It ran for 2340234 iterations. This is what it looked like after that:

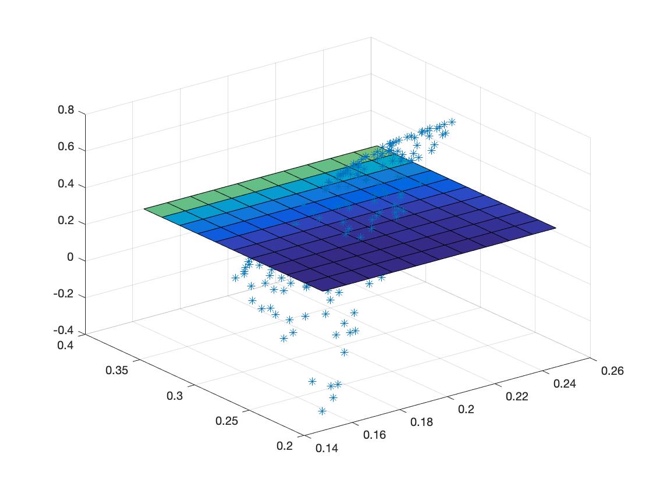
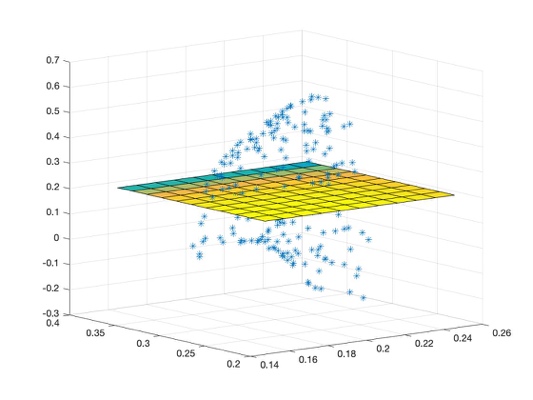


Next, I ran the preset network. As expected: this network converged much faster. This is the result after 5000 iterations.

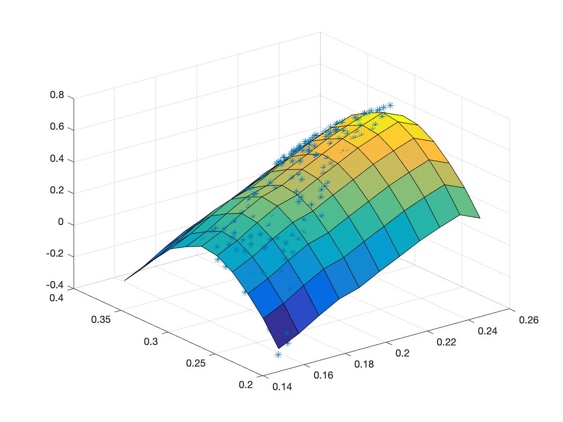
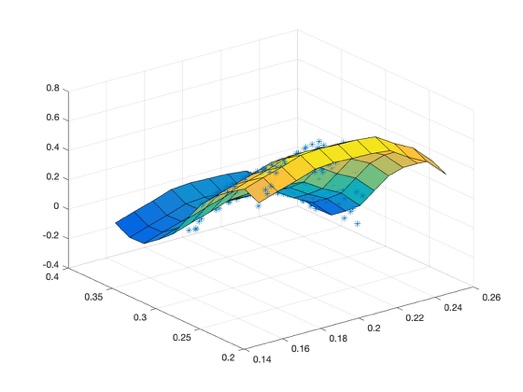


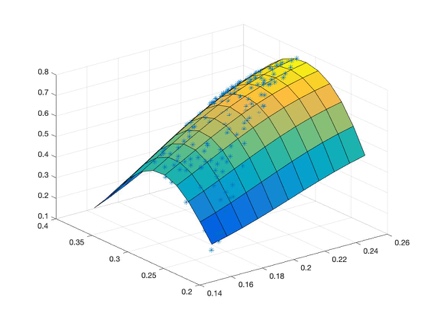
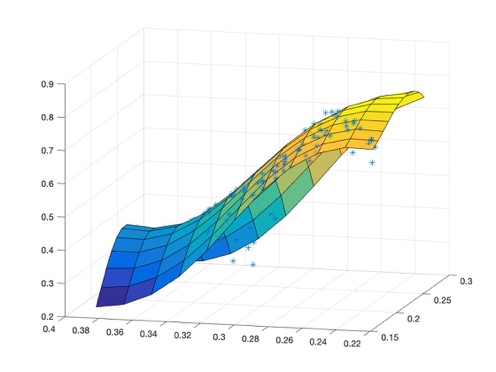


When scaling down the preset networks preset weights by a factor of 4 results in the network not being able to learn:



If we increase the number of interneurons in the final layer of the pretrained network to 100 the network converges after a few thousand iterations:



When the number of interneurons in the single hidden layer network is doubled to 40, the network converged on one mapping of the output extremely quickly, but took iterations to converge on the other: