Our NN will have a simple two-layer architecture. Input layer a[0]𝑎[0] will have 784 units corresponding to the 784 pixels in each 28x28 input image. A hidden layer a[1]𝑎[1] will have 10 units with ReLU activation, and finally our output layer a[2]𝑎[2] will have 10 units corresponding to the ten digit classes with softmax activation.

**Forward propagation**

Z[1]=W[1]X+b[1]𝑍[1]=𝑊[1]𝑋+𝑏[1]

A[1]=gReLU(Z[1]))𝐴[1]=𝑔ReLU(𝑍[1]))

Z[2]=W[2]A[1]+b[2]𝑍[2]=𝑊[2]𝐴[1]+𝑏[2]

A[2]=gsoftmax(Z[2])𝐴[2]=𝑔softmax(𝑍[2])

**Backward propagation**

dZ[2]=A[2]−Y𝑑𝑍[2]=𝐴[2]−𝑌

dW[2]=1mdZ[2]A[1]T𝑑𝑊[2]=1𝑚𝑑𝑍[2]𝐴[1]𝑇

dB[2]=1mΣdZ[2]𝑑𝐵[2]=1𝑚Σ𝑑𝑍[2]

dZ[1]=W[2]TdZ[2].∗g[1]′(z[1])𝑑𝑍[1]=𝑊[2]𝑇𝑑𝑍[2].∗𝑔[1]′(𝑧[1])

dW[1]=1mdZ[1]A[0]T𝑑𝑊[1]=1𝑚𝑑𝑍[1]𝐴[0]𝑇

dB[1]=1mΣdZ[1]𝑑𝐵[1]=1𝑚Σ𝑑𝑍[1]

**Parameter updates**

W[2]:=W[2]−αdW[2]𝑊[2]:=𝑊[2]−𝛼𝑑𝑊[2]

b[2]:=b[2]−αdb[2]𝑏[2]:=𝑏[2]−𝛼𝑑𝑏[2]

W[1]:=W[1]−αdW[1]𝑊[1]:=𝑊[1]−𝛼𝑑𝑊[1]

b[1]:=b[1]−αdb[1]𝑏[1]:=𝑏[1]−𝛼𝑑𝑏[1]

**Vars and shapes**

Forward prop

* A[0]=X𝐴[0]=𝑋: 784 x m
* Z[1]∼A[1]𝑍[1]∼𝐴[1]: 10 x m
* W[1]𝑊[1]: 10 x 784 (as W[1]A[0]∼Z[1]𝑊[1]𝐴[0]∼𝑍[1])
* B[1]𝐵[1]: 10 x 1
* Z[2]∼A[2]𝑍[2]∼𝐴[2]: 10 x m
* W[1]𝑊[1]: 10 x 10 (as W[2]A[1]∼Z[2]𝑊[2]𝐴[1]∼𝑍[2])
* B[2]𝐵[2]: 10 x 1

Backprop

* dZ[2]𝑑𝑍[2]: 10 x m ( A[2] 𝐴[2])
* dW[2]𝑑𝑊[2]: 10 x 10
* dB[2]𝑑𝐵[2]: 10 x 1
* dZ[1]𝑑𝑍[1]: 10 x m ( A[1] 𝐴[1])
* dW[1]𝑑𝑊[1]: 10 x 10
* dB[1]𝑑𝐵[1]: 10 x 1