1 Conventions

X is a $N \times D$ matrix (N being the number of samples and D the number of features) and W is a $D \times M$ matrix (M is the number of hidden units pr layer).

The output from the hidden layers is

$$Z = f\left(XW + b\right)$$

where f is an arbitrary mapping (i.e., sigmoid, relu) and b is the bias, a $M \times 1$ vector.

For some equations or examples, one may see the above equation in the form of

$$z = f(w^T x + b)$$

in which case, x is to be taken as a vector of features for a given sample, $D \times 1$, w is still a $D \times M$ matrix, and the bias vector b has the dimensions $M \times 1$.