zero gradients to propagate. Figure 4-7 illustrates sigmoidal and ReLU activations side by side.

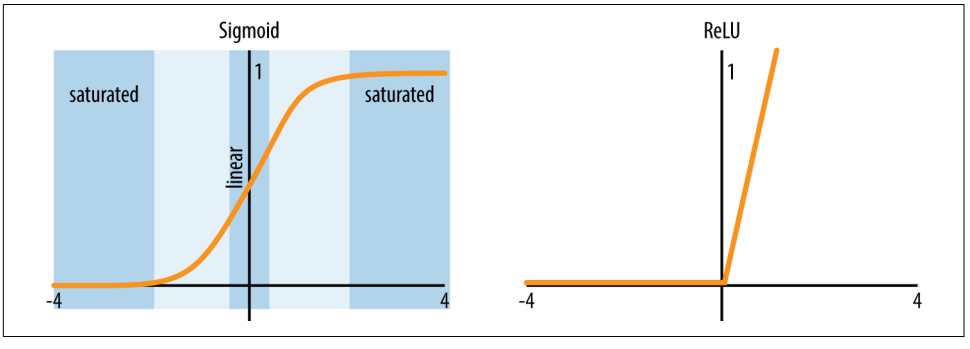

Figure 4-7. Sigmoidal and ReLU activation functions.

Fully Connected Networks Memorize

One of the striking aspects about fully connected networks is that they tend to memorize training data entirely given enough time. As a result, training a fully connected network to "convergence" isn't really a meaningful metric. The network will keep training and learning as long as the user is willing to wait.

For large enough networks, it is quite common for training loss to trend all the way to zero. This empirical observation is one the most practical demonstrations of the universal approximation capabilities of fully connected networks. Note however, that training loss trending to zero does not mean that the network has learned a more powerful model. It's rather likely that the model has started to memorize peculiarities of the training set that aren't applicable to any other datapoints.

It's worth digging into what we mean by peculiarities here. One of the interesting properties of high-dimensional statistics is that given a large enough dataset, there will be plenty of spurious correlations and patterns available for the picking. In practice, fully connected networks are entirely capable of finding and utilizing these spurious correlations. Controlling networks and preventing them from misbehaving in this fashion is critical for modeling success.

Regularization

Regularization is the general statistical term for a mathematical operation that limits memorization while promoting generalizable learning. There are many different types of regularization available, which we will cover in the next few sections.