## **Why do we need an adaptive learning rate?**

Why do we need an adaptive learning rate for every feature?

1. Consider input data with 4 features being processed through a sigmoid neuron
2. Here
3. From our gradient formula, we know that the value of the input feature plays a role in the gradient calculation i.e.
4. In real world scenarios, many features in the data are **sparse**, i.e. they take on a 0 value for most of the training inputs. Therefore, the derivatives corresponding to these 0 valued points are also 0, and the weight update is going to be 0.
5. To aid these sparse features, a larger learning rate can be applied to the **non-zero valued points** of these sparse features.
6. Example:
   1. Consider a subject at college that you are taught for only 5 minutes a day
   2. For those 5 minutes, maximising your attention span would allow for maximum knowledge retention
   3. In this case, the 5-minute subject would be a sparse feature i.e. a feature that does not occur very often in the training data
   4. And the attention span would be our learning rate. A high learning rate for the sparse features allows us to maximise the learning (weight updation) we get from it.
7. Conversely, **dense** features are those with non-zero values for most of the data points. They must be dealt with by using a lower learning rate.
8. Can we have a different learning rate for each parameter(weights) which takes care of the frequency(sparsity/density) of features?