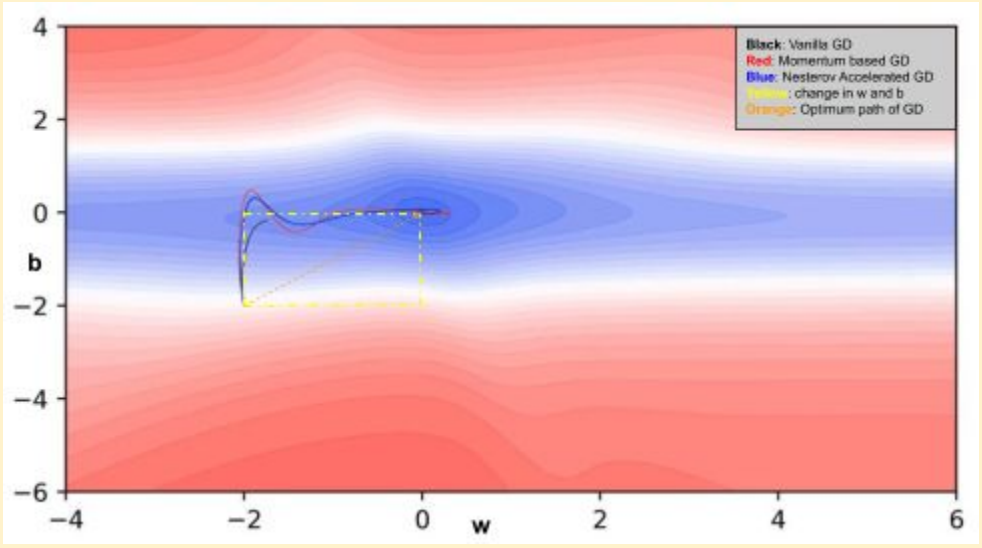
PadhAI: Variants of Gradient Descent

One Fourth Labs

Running and Visualizing Adagrad

Let's compare this to vanilla, momentum based, NAG gradient descent

1. Let's plot the 2D visualisation of vanilla, momentum based, NAG gradient descent



2. Here, w & b behave as two features of the input (x_0, x_1). b is a dense feature and is always a non-zero value. w is deliberately chosen as a sparse feature with 80% of the values as 0.
3. Thus, we would need a higher learning rate for w and a lower learning rate for b, if not, we will end up with sub-optimal paths as shown by the previous 3 types of GD from the figure.
4. Let's look at a visualisation of Adagrad