1. Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.
   Network is initialized with all weights as zero.
   1.1 Output is zero.
   1.2 All derivatives ($\frac{\partial L}{\partial w}$) are zero.
   1.3 Loss is zero.
   1.4 With Back propagation, weights won't change.
   1.5 None of the above.

   ABD

2. Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.
   Network is initialized with all weights as non-zero but a small constant.
   2.1 Output is zero.
   2.2 All derivatives ($\frac{\partial L}{\partial w}$) are zero.
   2.3 Loss is zero.
   2.4 With Back propagation, weights won't change.
   2.5 None of the above.

E

3. Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

   What may be a better initialization, among the following?

   3.1 All weights the same.
   3.2 All the weights random and positive
   3.3 All the weights random and negative.
   3.4 Some weights random and positive and some weights random and negative.
   3.5 All of the above are equally good or equally bad.

   BCD

4. Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

   We use this network for regression to predict, say the mean temperature of tomorrow in Hyderabad, which is always positive.

   We train the network with sufficient amount of data, and follow good practices of training.

4.1 At the end of training, we are at a local minima.
4.2 At the end of training, we will be at a global minima.
4.3 At the end of training Loss will become zero.
4.4 At the end of training, all our weights will be positive.
4.5 None of the above.

A

5. Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.
We initialize the network with good practices available/reported. Starting from the same initialization, we train the network multiple times with BP/GD. Starting from the same initialization, we train the network multiple times with BP/GD with momentum term also.
   5.1 Starting from the same initialization, the solution at epoch 100 remains same in all the runs, when the implementation was SGD.
   5.2 Starting from the same initialization, the solution at epoch 100 remains same in all the runs, when the implementation was batch GD.

5.3 Starting from the same initialization, the solution at epoch 100 remains same with or without momentum.

5.4 With an appropriate but fixed convergence criteria (say early stopping), models trained with and without momentum will be the same.

5.5 With an appropriate but fixed convergence criteria (say early stopping), models trained with and without momentum could be different.

BE the model without momentum is more likely to be in local minima. In GD, the gradient is from all samples; so will always be same in both runs whereas in SGD the sample are chosen in random, so the gradients are different