How did you test whether your code gives the correct output? ¶

As seen in each of the error vs iteration plot the train error has saturated (except for Adagrad which is expected to perform poor). Also the test error is in simalar range of train error, thus suggesting it performs well.

Write briefly what you think the relative merits and demerits of this variant are.

- In SGD each update step is cheaper than Batch Gradient Descent. Also has the ability to escape local minima and saddle points. Will reach minimum if run for certain epochs.
- SGD with momentum reduces the oscillations i.e movement perpendicular to direction of minima, by accelerating if 2 consecutive updates are in same direction and retarding otherwise. Thus leading to faster convergence also can be seen in the plot.
- Adagrad was formed to be used with sparse data i.e where gradient updates for different variables may require to be of different order. So a sparse data may require a bigger update than frequently occurring data. Performs poorly here since learning rate is attenuated very quickly.
- RMSprop does something similar to Adagrad but doesn't penalise frequent data as much. It still attenuates but in a weighted fashion.