The dataset we are using is MNIST with 60000 training samples and 10000 testing samples. The task is to classify 28*28-pixel images of hand-written numbers into its corresponding number.

The experiments are done on a fully-connected neural network with 2 hidden layers, each of which in size 800, 500, with output size 10, one for each class. We simulate the distributed procedure with N threads, each representing a worker, among them s are faulty agents that would calculate the correct gradient, reverse it and time 100. For each epoch, there would be s workers randomly selected to be faulty.

The 60000 training samples are evenly distributed to N workers in the beginning to simulate the situation where different workers have training samples of their own. Each step every worker will calculate the average gradient of a b-size batch of samples from their own $60000/N$ samples.

In the following experiments, set $N = 40$, $s = 2$, $b = 128$. Each worker gets $60000/40 = 1500$ samples, and each epoch contains $1500/128 = 12$ batches, of which the last batch only contain 92 samples.

The filters that are used are averaging (normal), krum, median of means, coordinate-wise median, coordinate-wise trimmed mean and gradient norm clipping.

The loss function is cross entropy. The results are evaluated using precision@k, where if among the first k results there is the correct answer, the prediction is counted as true positive. Here $k = 1, 5$.

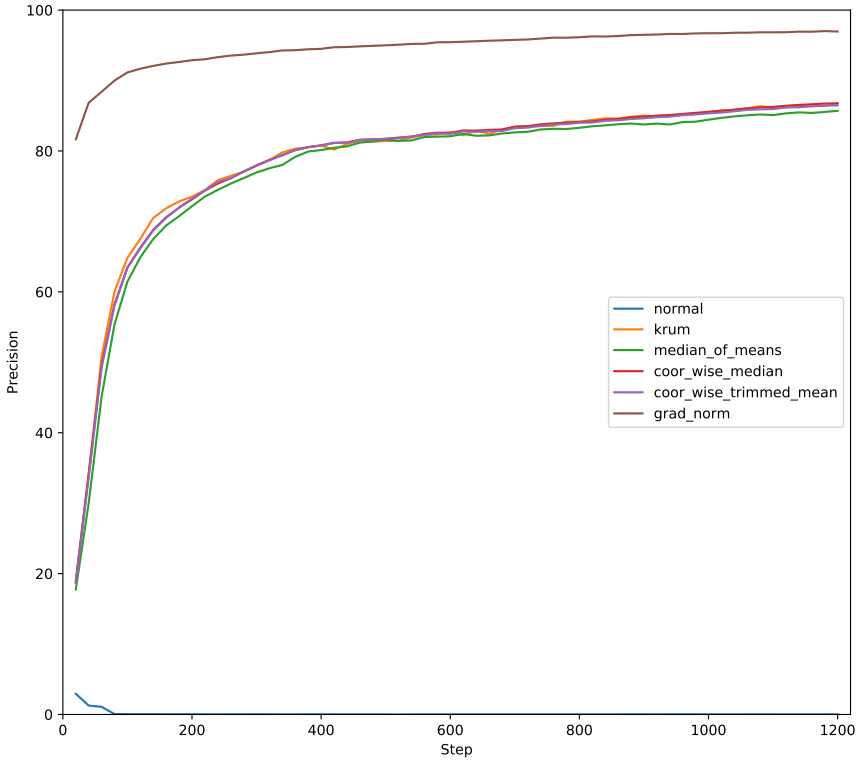

Figure 1: Precision@1

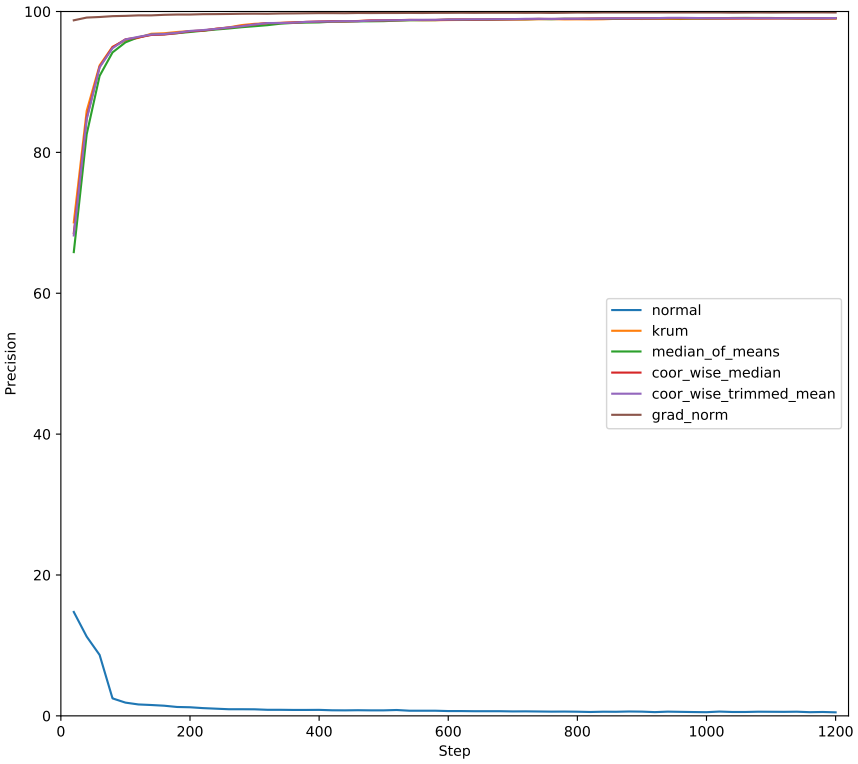

Figure 2: Precision@5