EP3260: Fundamentals of Machine Learning over Networks
Computer Assignment 4, Due Date: March 27, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Computer Assignment 4. *(Sensitivity to Outliers)*
Split "MNIST" dataset to 10 random disjoint subsets, each for one worker, and consider SVM classifier in the form of $\min_{\boldsymbol{w}} \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w})$ with $N = 10$. Consider the following outlier model: each worker i at every iteration independently and randomly with probability p adds a zero-mean Gaussian noise with a large variance R to the information it shares, i.e., ∇f_i and $\boldsymbol{w}_{j,k}$ in the cases of Algorithm 1 and decentralized subgradient method of Lecture 6, respectively.

(a) Run decentralized gradient descent (Algorithm 1) with 10 workers.

- Characterize the convergence against p and R.

Solution. In this part, we run a decentralized algorithm using 10 workers. In the following figures, we illustrate the loss versus the number of iteration for different values of probabilities and variances. We can see that the loss converges for some values of probability and variance. The loss function for x_i is computed using $L_i = \sum_{j\neq y_i} \max(0, (w_j - w_{y_i})^T x_i + \delta)$. We minimize the average loss using decentralized gradient descent.

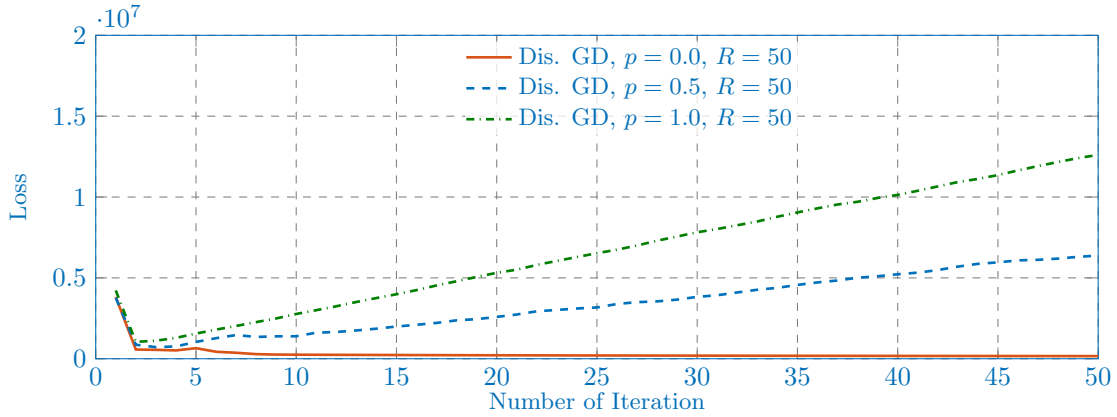


Figure 1: Loss versus iteration for fixed $R = 50$.

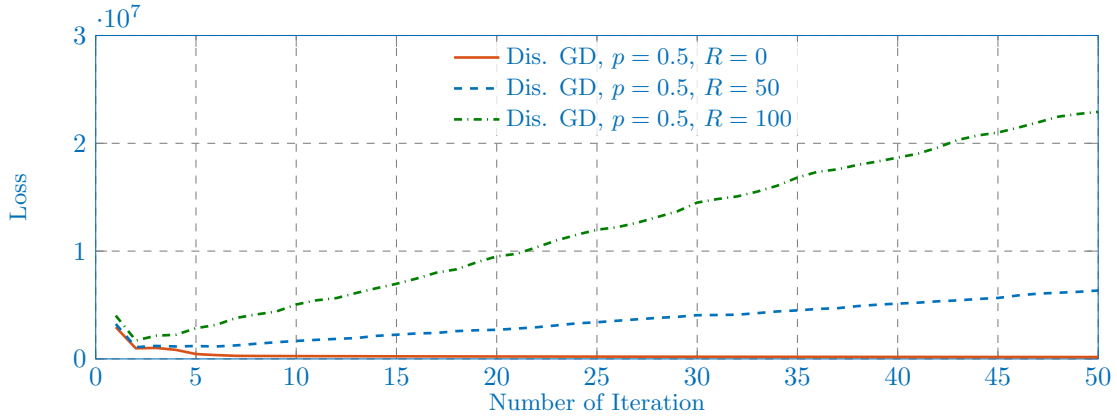


Figure 2: Loss versus iteration for fixed $p = 0.5$.

- Propose an efficient approach to improve the robustness of Algorithm 1 and characterize its convergence against p and R.

Solution. The master node updates the gradient with respect to $\boldsymbol{w}_k$ using a weighted average of the gradients the workers. The weight for a worker depends on the similarity of its gradient to the ones for the rest of workers. Here, we used the cosine operator for similarity. Also, we propose to transmit workers' messages to the master node multiple times, so that the master can average the received gradients and obtain more accurate vlaues. The following figures show the results of running the robust decentralized gradient descent algorithm.

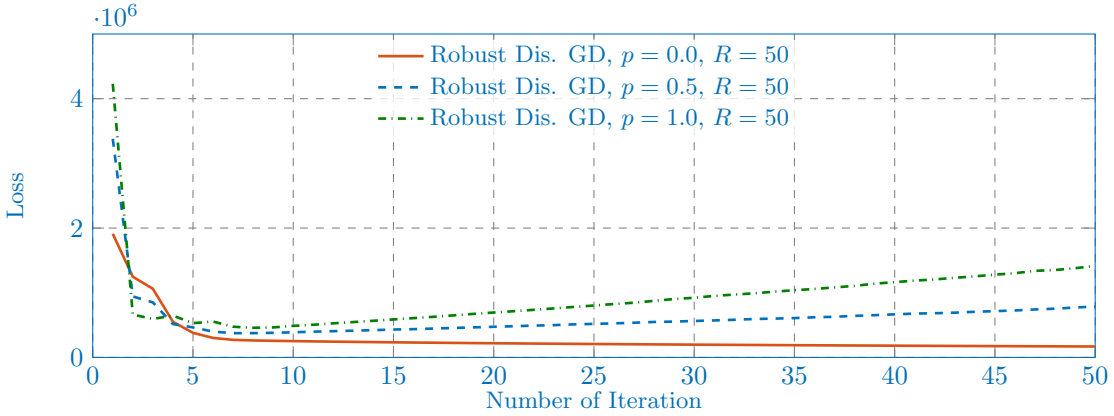


Figure 3: Loss versus iteration for fixed $R = 50$.

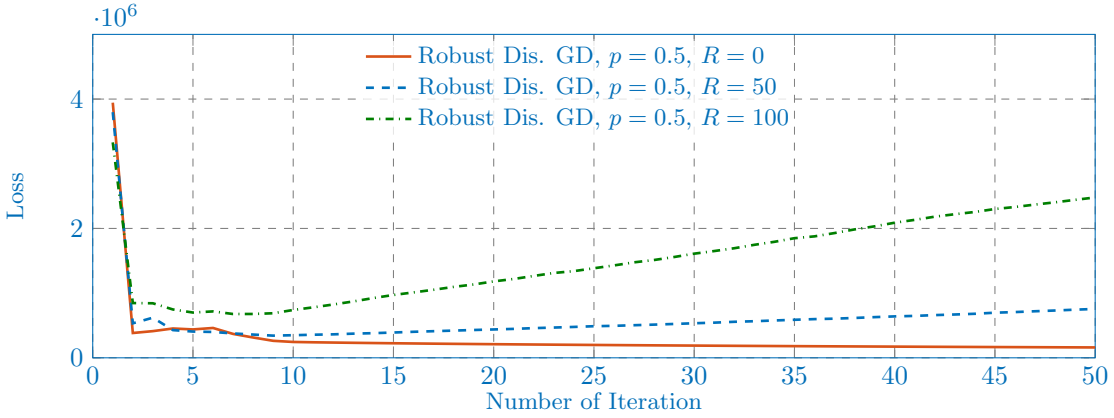


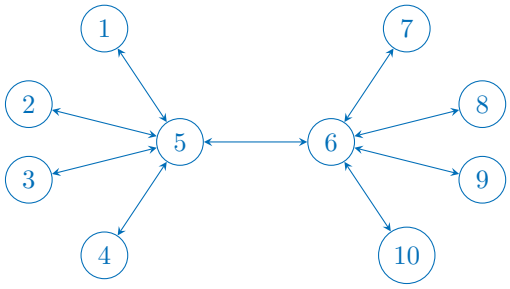
Figure 4: Loss versus iteration for fixed $p = 0.5$.

It can be seen through above figures, the algorithm is more robust to noise. □

(b) Consider a two-star topology with communication graph $(1, 2, 3, 4)$-5-6-$(7, 8, 9, 10)$ and run decentralized subgradient method.

- Characterize the convergence against p and R.

Solution. The following figure shows the topology of the network. For this part, we calculated the average consensus value. In the following figures, the loss converges to a fixed level for different values of probabilities and variances.



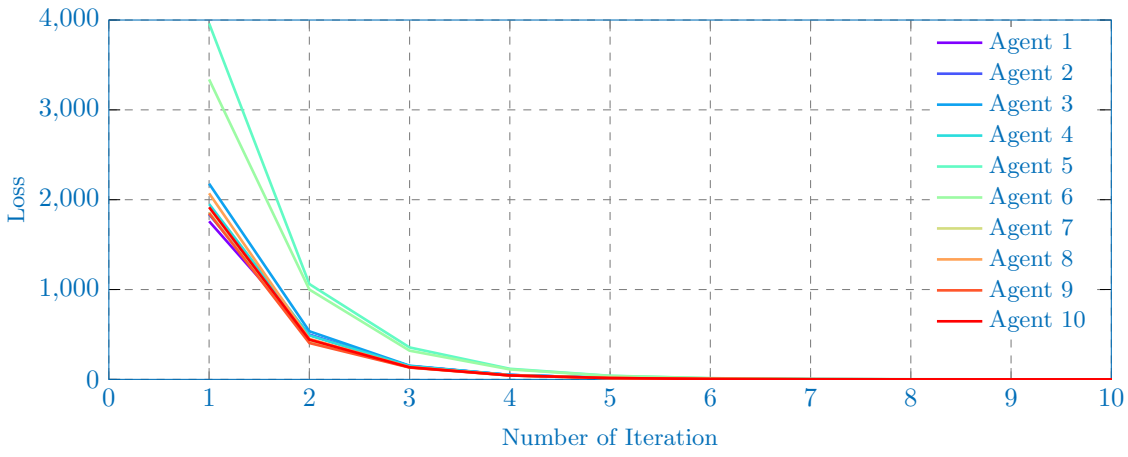


Figure 5: Loss versus iteration for distributed SM, $p = 0.0$, $R = 0.01$.

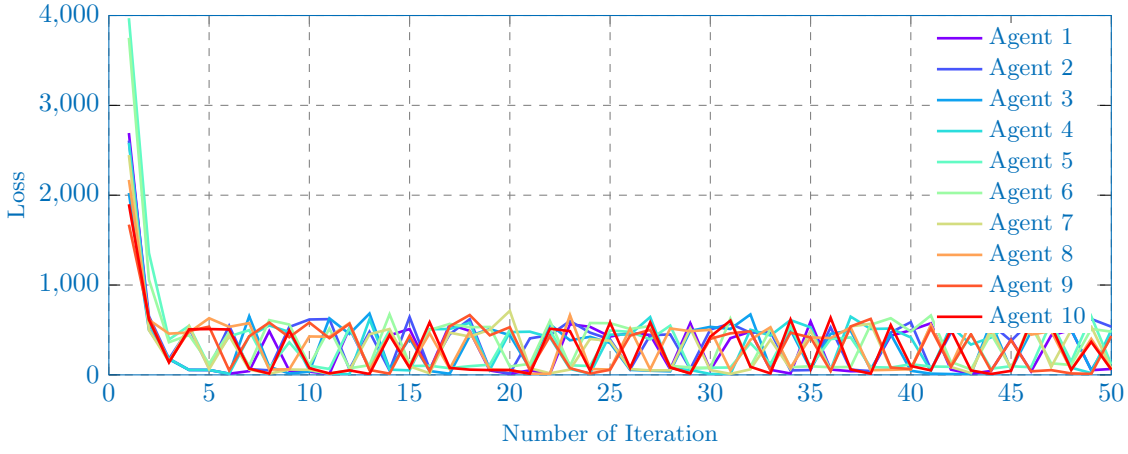


Figure 6: Loss versus iteration for distributed SM, $p = 0.5$, $R = 0.01$.

□

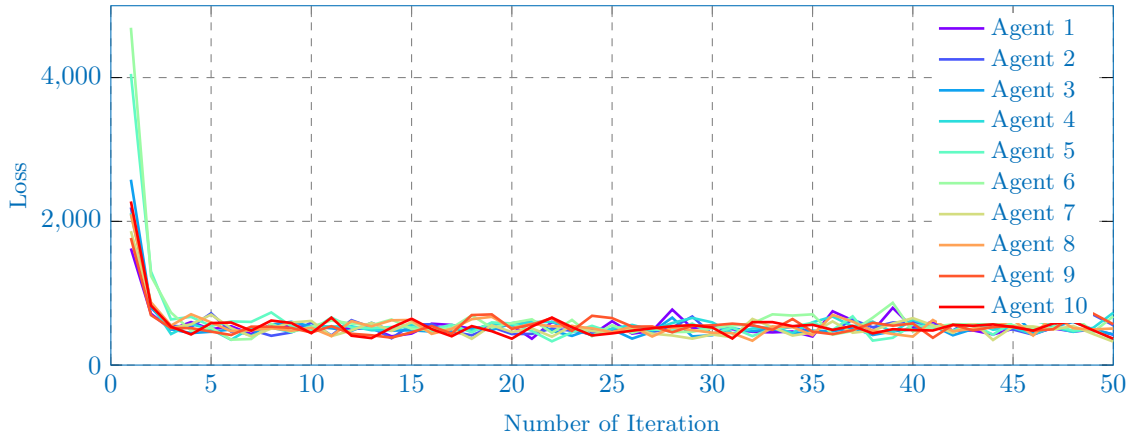


Figure 7: Loss versus iteration for distributed SM, $p = 1$, $R = 0.01$.

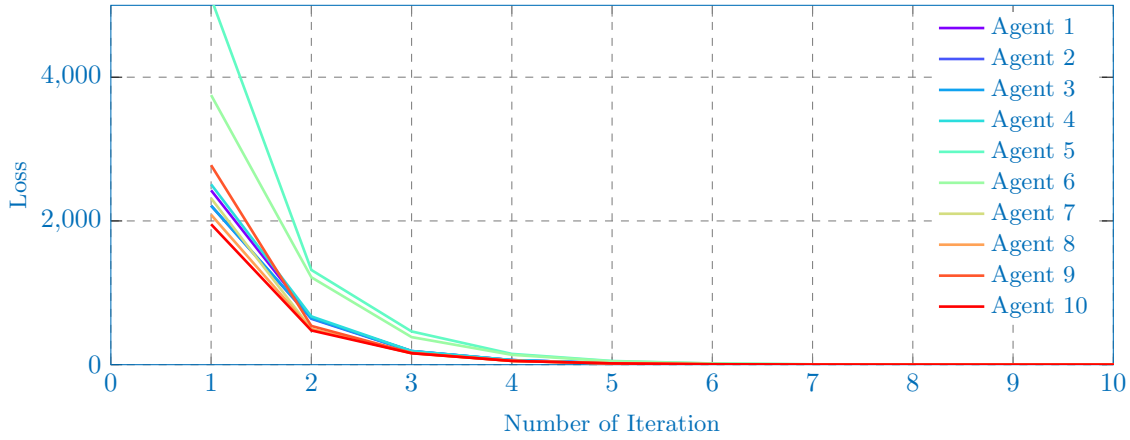


Figure 8: Loss versus iteration for distributed SM, $p = 0.5$, $R = 0$.

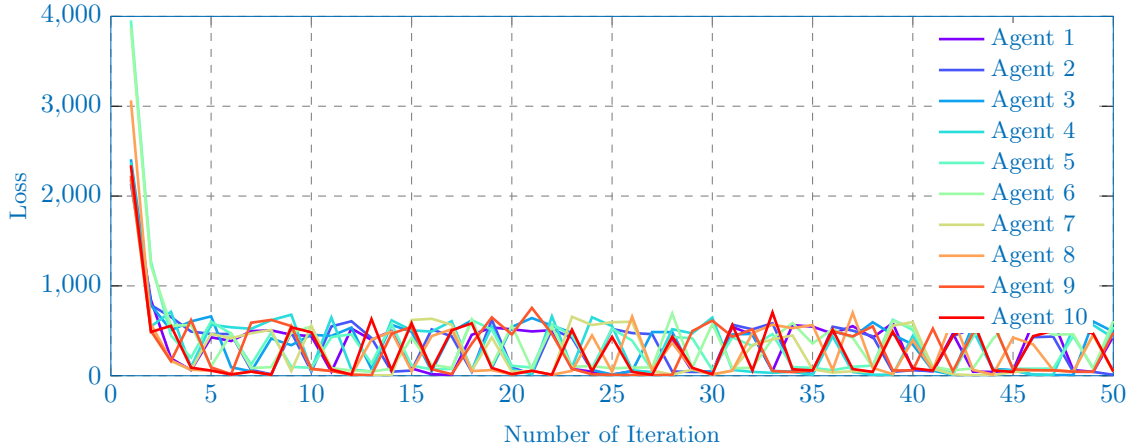


Figure 9: Loss versus iteration for distributed SM, $p = 0.5$, $R = 0.01$.

- Propose an efficient approach to improve the robustness to outliers and characterize its convergence against p and R.

  Solution. We use multiple transmissions, to have multiple instances of noisy gradients from each worker. Averaging these instances results in a more accurate estimate of the gradient compared to the single transmission. Therefore the algorithm will be more robust to the outliers. The following figures demonstrate the results of the robust algorithm.

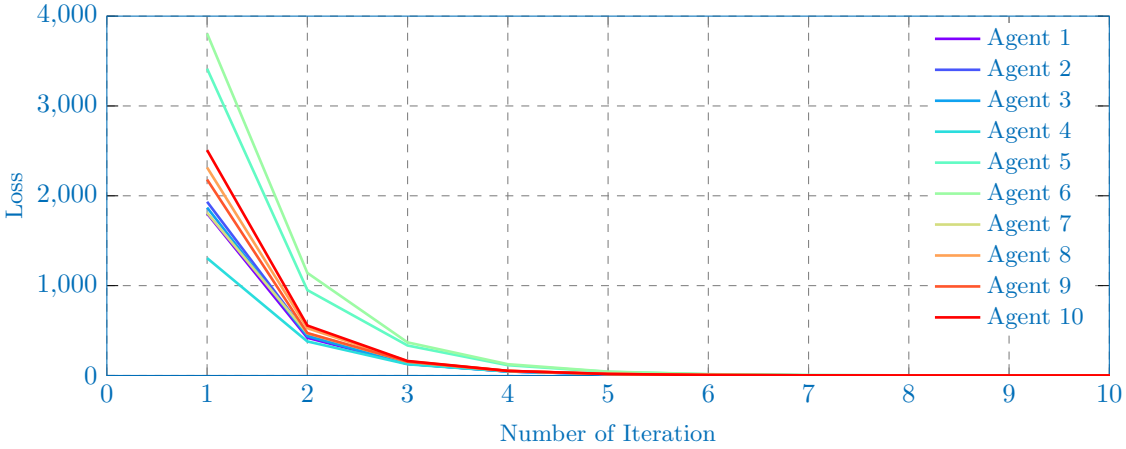


Figure 10: Loss versus iteration for robust distributed SM, $p = 0.0$, $R = 0.01$.

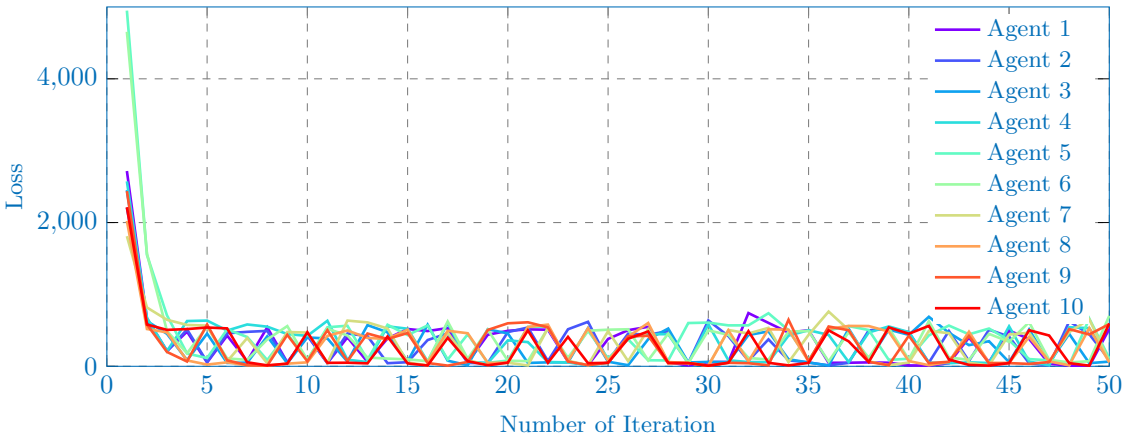


Figure 11: Loss versus iteration for robust distributed SM, $p = 0.5$, $R = 0.01$.

□

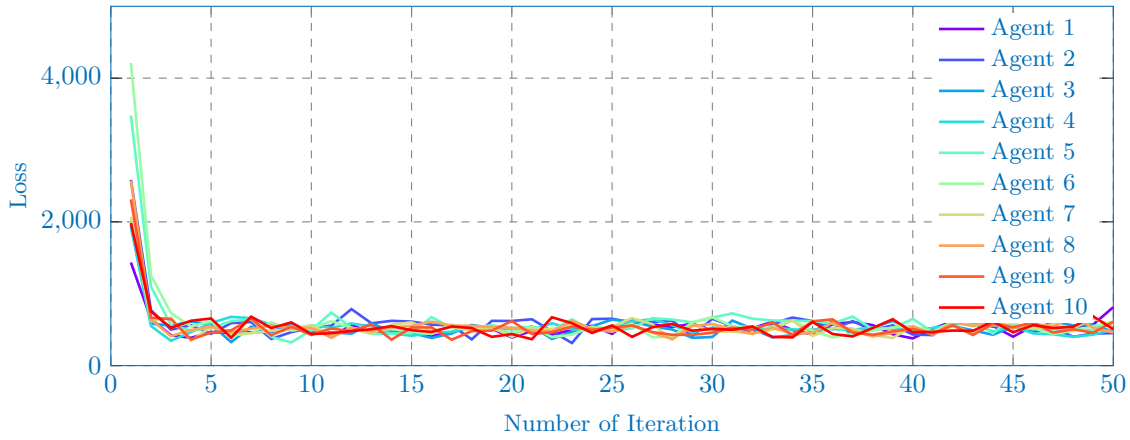


Figure 12: Loss versus iteration for robust distributed SM, $p = 1$, $R = 0.01$.

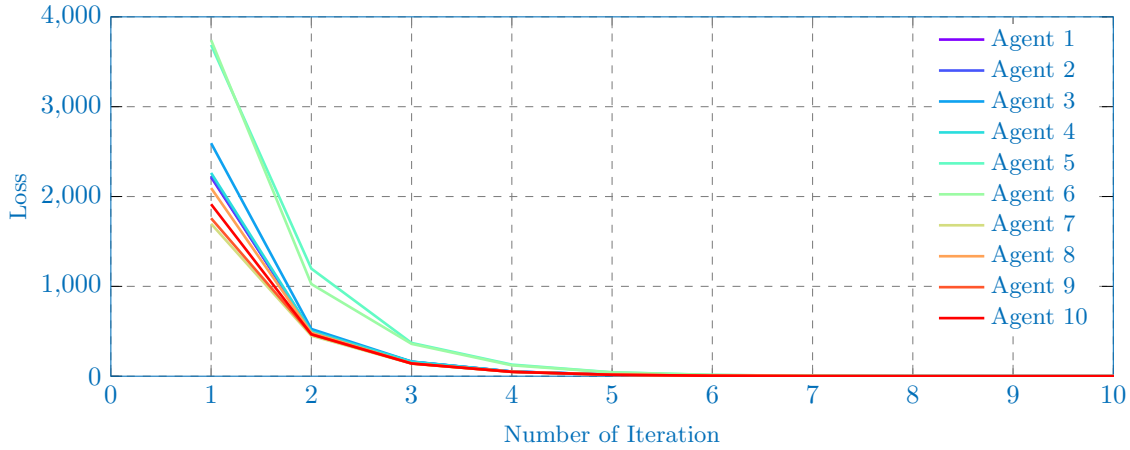


Figure 13: Loss versus iteration for robust distributed SM, $p = 0.5$, $R = 0.0$.

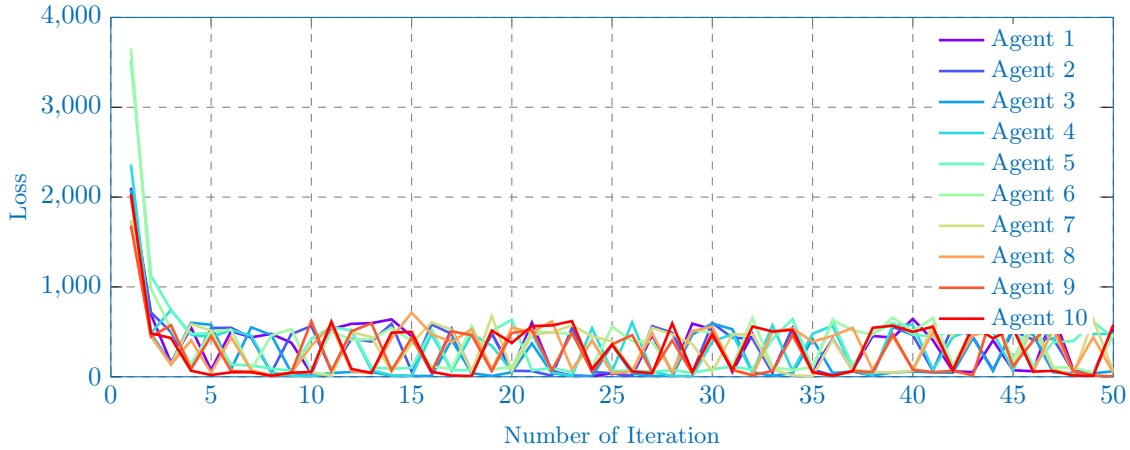


Figure 14: Loss versus iteration for robust distributed SM, $p = 0.5$, $R = 0.01$.

(c) Assume that we can protect only three workers in the sense that they would always send the true information. Which workers you protect in Algorithm 1 and which in the two-star topology, running decentralized subgradient method?

Solution. In Algorithm 1, the master node should be protected. Moreover, two other nodes that are more similar to the rest of the nodes should be protected. In other words, the two nodes that have more weight should be protected.

In the two-star topology, nodes 5 and 6 (i.e., star nodes) should be protected. Also, one more node can be chosen. For calculation of the average in the robust method, we should assign more weight to the nodes that are protected.

□