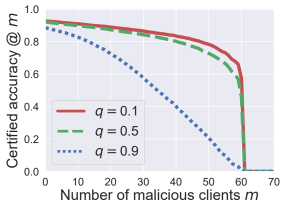

Figure 6: Impact of the degree of non-IID q on MNIST.

Impact of k, N, and α: Figure **??**, **??**, and **??** show the impact of k, N, and α, respectively. k achieves a trade-off between accuracy under no malicious clients and security under malicious clients. Specifically, when k is larger, the ensemble global model is more accurate at $m = 0$, but the certified accuracy drops more quickly to 0 as m increases. This is because when k is larger, it is more likely for the sampled k clients to include malicious ones. The certified accuracy increases as N or α increases. This is because training more global models or a larger α allows Algorithm **??** to estimate tighter probability bounds and larger certified security levels. When N increases from 100 to 500, the certified accuracy increases significantly. However, when N further grows to 1,000, the increase of certified accuracy is marginal. Our results show that we don't need to train too many global models in practice, as the certified accuracy saturates when N is larger than some threshold.

Impact of degree of non-IID q: Figure **??** shows the certified accuracy of our ensemble FedAvg on MNIST when the clients' local training data have different degrees of non-IID. We observe that the certified accuracy drops when q increases from 0.5 to 0.9, which represents a high degree of non-IID. However, the certified accuracy is still high when m is small for $q = 0.9$, e.g., the certified accuracy is still 83% when $m = 10$. This is because although each global model trained using a subsample of clients is less accurate when the local training data are highly non-IID, the ensemble of multiple global models is still accurate.

Related Work

In federated learning, the first category of studies (**?????**) aim to design federated learning methods that can learn more accurate global models and/or analyze their convergence properties. For instance, FedMA (**?**) constructs the global model via matching and averaging the hidden elements in a neural network with similar feature extraction signatures. The second category of studies (**????????????????**) aim to improve the communication efficiency between the clients and server via sparsification, quantization, and/or encoding of the model updates sent from the clients to the server. The third category of studies (**????????**) aim to explore the privacy/fairness issues of federated learning and their defenses.These studies often assume a single global model is shared among the clients. Smith et al. (**?**) proposed to learn a customized model for each client via multi-task learning.

Our work is on security of federated learning, which is orthogonal to the studies above. Multiple studies (**????**) showed that the global model's accuracy can be significantly downgraded by malicious clients. Existing defenses against malicious clients leverage Byzantine-robust aggregation rules such as Krum (**?**), trimmed mean (**?**), coordinate-wise median (**?**), and Bulyan (**?**). However, they cannot provably guarantee that the global model's predicted label for a testing example is not affected by malicious clients. As a result, they may be broken by strong attacks that carefully craft the model updates sent from the malicious clients to the server, e.g., (**?**). We propose ensemble federated learning whose predicted label for a testing example is provably not affected by a bounded number of malicious clients.

We note that ensemble methods were also proposed as provably secure defenses (e.g., (**?**)) against data poisoning attacks. However, they are insufficient to defend against malicious clients that can manipulate both the local training data and the model updates. In particular, a provably secure defense against data poisoning attacks guarantees that the label predicted for a testing example is unaffected by a bounded number of poisoned training examples. However, a single malicious client can poison an arbitrary number of its local training examples, breaking the assumption of provably secure defenses against data poisoning attacks.

Conclusion

In this work, we propose ensemble federated learning and derive its tight provable security guarantee against malicious clients. Moreover, we propose an algorithm to compute the certified security levels. Our empirical results on two datasets show that our ensemble federated learning can effectively defend against malicious clients with provable security guarantees. Interesting future work includes estimating the probability bounds deterministically and considering the internal structure of a base federated learning algorithm to further improve our provable security guarantees.

Acknowledgement

We thank the anonymous reviewers for insightful reviews. This work was supported by NSF grant No.1937786.

Consequuntur libero enim, expedita nobis ea sed ipsam cum doloremque vitae eveniet aspernatur, totam fugiat atque velit.Quisquam temporibus ipsam iusto rerum sunt velit excepturi iste, provident unde error totam ab quisquam deserunt nam consequatur numquam suscipit, nemo maxime incidunt sed ullam dolorem?Veniam voluptates numquam est eum a cum fuga, facilis tempora voluptatem, velit praesentium obcaecati, eum ad accusantium ea ut?Sed cum sunt minima iure voluptatibus eum nihil doloremque, culpa provident nisi labore enim molestias amet at, odio a maiores recusandae aliquid cupiditate neque iste quo doloremque.Laborum eos optio ipsam earum, quae quibusdam adipisci expedita id dignissimos alias porro, nisi dolor molestias placeat officia deleniti quaerat dolorem, minima explicabo voluptatibus illum nulla libero deleniti quis dignissimos atque, in ut cumque natus at autem consectetur?Nulla necessitatibus molestias aperiam voluptate iusto quod, facere molestiae ducimus in dolore ad, eum