Defending Distributed Systems Against Adversarial Attacks: Consensus, Consensus-based Learning, and Statistical Learning

Lili Su

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology

Homepage: https://sites.google.com/site/lilisuece/

Thesis: Defending Distributed Systems Against Adversarial Attacks: Consensus, Consensus-based Learning, and Statistical Learning

Advisor: Prof. Nitin H. Vaidya, Georgetown University (previously at UIUC)

Brief Biography: I am a postdoc in the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT, hosted by Professor Nancy Lynch. She received a Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2017, supervised by Professor Nitin H. Vaidya. Her research intersects distributed systems, learning, security, and brain computing. She was the runner-up for the Best Student Paper Award at DISC 2016, and she received the 2015 Best Student Paper Award at SSS 2015. She received UIUC's Sundaram Seshu International Student Fellowship for 2016, and was invited to participate in Rising Stars in EECS (2018). She has served on TPC for several conferences including ICDCS and ICDCN.

Research Summary: Distributed systems are ubiquitous both in human society and in nature. In human society, due to constraints in computation power and data accessibility, many learning systems are distributed, such as Federated Learning (FL), Internet of Things (IoT), and multi-agent networks. In nature, neural circuits in the brain and social insect colonies are biological distributed systems. In fact, the biological distributed systems and their dynamics have inspired many influential artificial counterparts which include deep neural networks and ant colony optimization algorithms.

I have explored topics ranging from adversary-resilience in Federated Learning and multi-agent networks, to decision-making in social insect colonies, to neural networks training, and to neural computation. My plan for the near future is to continue this broad exploration with primary focuses on developing provably secured algorithms for unreliable learning systems such as FL, and on understanding (both artificial and biological) neural networks. Nevertheless, I see myself pursuing new directions as technology evolves and opportunities arise. In the long run, I plan to employ theoretically-grounded principles to design efficient methods for attacking real-world problems in *artificial* distributed systems, and to provide testable hypotheses on the underlying mechanisms of some observable behaviors or phenomena of *biological* distributed systems.

I briefly describe the problems and summarize our results on adversary-resilience in Federated Learning, adversary-resilience in multi-agent networks, neural networks and neural computation, and biologically-inspired distributed algorithms, respectively.

1. Adversary-Resilience in Federated Learning.

With the rapid advancements in data collection, storage, and computation capabilities of smartphones, and with the growing popularity of wearable devices such as Apple Watch, one trend in machine learning is to "outsource" part of the computation burden to edge and/or end devices. This trend is motivated by not only computation and response speed benefits but also privacy gains. In view of this trend, Google proposes *Federated Learning* – a new learning paradigm [1, 3]. Compared with traditional learning, FL suffers serious security issues [5, 3] and several practical constraints call for new security strategies. For example, since the local data volume is typically low in FL, it is hard for the cloud (i.e., the leaner) to distinguish errors injected by malicious workers from random errors induced by honest workers based on their message values only. Besides, the communication between the cloud and the end devices suffer low-throughput and high-latency.

To the best of our knowledge, we are the first to study Byzantine-resilience in FL [5], which has been attracting more and more research attention [2, 24, 25, 10, 8]. Concurrent with our work [5], a similar problem was proposed in [4], which, in sharp contrast to FL, considered the setting wherein all workers operate on a common dataset. This difference is fundamental as can be seen from the analysis difference in [4] and [5]. In [5], we designed a Byzantine-resilient gradient descent method in which the learner uses *geometric median of means* to aggregate the gradients reported by the workers. We showed that when the number of Byzantine workers is sufficiently small, with high probability, the ℓ_2 error converges in $O(\log N)$ rounds to an estimation error $O(\sqrt{dq/N})$, where d is the model dimension, q is the upper bound of the tolerable Byzantine workers, and N is the total number of data points collectively kept by all workers. In a follow-up work by ourselves [19], under the same set of technical assumptions, we proposed another variant of gradient descent in which, within each round of gradient descent update, the learner adopts an *iterative filtering* rule to aggregate the gradients. The ℓ_2 error converges in $O(\log N)$ rounds to $O\left(\sqrt{q/N} + \sqrt{d/N}\right)$, matching the op-

timal error rate $O(\sqrt{d/N})$ when $q = O(d)$.

On the technical front, a key challenge is that Byzantine failures might create arbitrary and unspecified dependency among the gradient descent updates at different rounds. To handle this issue, in both [5] and [19], we proved that the aggregated gradient, as a function of model parameter, converges *uniformly* to the true gradient function. In addition, in [19], deviating from the existing literature on robustly estimating a finite-dimensional mean vector, we establish a uniform concentration of the sample covariance matrix of gradients. To get a near-optimal uniform concentration bound, we develop a new matrix concentration inequality, which might be of independent interest.

Significant progress has been made [2, 24, 25, 19, 10, 8]. Nevertheless, many interesting and practically important problems remain unsolved. I plan to continue working intensively in this area in the future.

2. Adversary-Resilience in Multi-Agent Networks.

Many multi-agent networks such as IoT and micro-grids are also vulnerable to unstructured faults. This is because in large distributed systems, due to unreliable devices and communication channels, and even external adversarial attacks, individual computing devices/sensors might exhibit abnormal behaviors. Such abnormal behaviors are often unstructured because of the heterogeneity in hardwares, softwares, implementation environments, and the unpredictability of external adversarial attacks. Adversary-resilient distributed computing dates back to the *Byzantine General Problem* in computer science [9], and is getting more and more research attention in communication and control communities [6, 22]. In contrast to fault-free networks, dealing with Byzantine faults is very challenging. It is well-known that in complete graphs, no consensus algorithms can tolerate more than 1/3 of the agents to be Byzantine [9]. In addition, Byzantine consensus with multi-dimensional inputs in the complete graphs had not been solved until recently [11, 23].

(1) Multi-Agent Optimization: In [15, 17] and a series of technical reports, we considered the problem of multi-agent optimization wherein an unknown subset of agents suffer Byzantine faults. Specifically, each agent i has a local cost function f_i, and the overarching goal of the good agents is to collaboratively minimize a global objective that properly aggregates these local cost functions. To the best of our knowledge, we are among the first to study Byzantine-resilient optimization where no central coordinating agent exists, and we are the first to characterize the structures of the convex coefficients of the achievable global objectives. Recently, we organized the results scattered in our unpublished technical reports and the preliminary conference works [15, 17], and combined them into one writeup [16]. In contrast to [15, 17], in [16], we focused on the general networks which include complete networks as special cases.

(2) Non-Bayesian Learning: To avoid the complexity of Bayesian learning, an approximate Bayesian learning framework in networks, referred to as *Non-Bayesian Learning*, was proposed [7]. The prior work implicitly assumes that the networked agents are reliable in the sense that they correctly follow the specified distributed algorithm. However, in some practical multi-agent networks, this assumption may not hold. We proposed and analyzed a learning rule [18] wherein each agent updates its local pseudo beliefs as (up to normalization) the product of (A) the likelihood of the cumulative private signals and (B) the weighted geometric average of the beliefs of its incoming neighbors and itself (using Byzantine consensus). I was the runner-up for the **Best Student Paper Award at DISC 2016** for my work [18].

Under the above two themes, there are many interesting future directions such as the tradeoff between network redundancy and information redundancy, and communication/computation efficient algorithms. Beyond these two themes, there are many interesting applications such as distributed state estimation problem, as we studied in [14], and autocar security.

3. Neural Networks and Neural Computation.

Despite intensive research efforts, a thorough understanding of the theory behind the practical success of artificial neural networks, even for two-layer neural networks, is still lacking. For example, in sharp contrast to traditional learning theory, over-parameterized neural networks are observed to enjoy smaller training and even smaller generalization errors [26], i.e., they do not overfit. It has been shown that with proper random network initialization, (S)GD converges to a (nearly) global minimum provided that the width of the network is *polynomially* over-parameterized. However, neural networks seem to interpolate the training data as soon as the number of parameters exceeds the size of the training dataset by a constant factor [26]. In our recent work [20], we showed that nearly-linear over-parameterization is sufficient for two-layer network training. To the best of our knowledge, this is the first result showing the sufficiency of nearly-linear over-parameterization. Moreover, in contrast to existing convergence rates which approach 0 as the dataset size grows, we characterized a *constant* convergence rate (w. r. t. the size of the dataset). Such rate characterization is important as in many applications the dataset volumes are huge – the ImageNet dataset has 14 million images. Nevertheless, we only considered the setting wherein the dimension of the input feature is fixed, leaving the high dimensional region as one future direction.

In addition to artificial neural networks, I have also worked on neural computation in the brain [13]. We presented a framework for studying a noisy Winner-Take-All computation in a spiking neural network. Winner-Take-All (WTA) is a hypothesized mechanism in the brain to select proper neurons from a competitive network of neurons, and is conjectured to be a fundamental primitive of cognitive functions such as attention and object recognition. In our work, time is slotted into intervals of $1\,ms$. The inputs are modeled by independent Bernoulli processes in time with fixed rates; here Bernoulli processes are assumed in order to capture the fact that biological neurons have a refractory period, i.e., typically, a neuron cannot spike twice within $1\,ms$. Spiking neural networks are getting more and more research attention in computational neuroscience due to both its bio-relevance and its energy efficiency. In [13], we obtained an information-theoretic lower bound on the waiting-time to obtain a given accuracy, and constructed a simple neural circuit that turns out to be order-optimal for a given accuracy (fixed).

Both [20] and [13] are my first attempt towards understanding artificial and biological neural networks, respectively. In the future, I would like to further study artificial and biological neural networks, respectively, yet with the hope of drawing inspirations and insights from one another

to arrive at better understanding of both of them.

4. Biologically-Inspired Distributed Algorithms.

Social insect colonies like ants have been existed for about 140 million years. However, we human beings do not know much about ants. Understanding ant colonies is very challenging – the estimated number of ant species is 22,000, each of them could have different behaviors and underlying mechanisms in generating certain observable behaviors. Though each ant is simple, a collection of ants can collaboratively perform complex tasks such as house-hunting and task allocation.

In [21], we provided some insights for the underlying mechanism for house-hunting in ant colonies. House-hunting refers to the collective decision process in which the entire social group collectively identifies a high-quality site to immigrate to. For the success of house-hunting, individuals repeatedly scout and evaluate multiple candidate sites, and exchange information with each other to reach a collective decision. House-hunting can be viewed as sequential decision problems, and can be naturally modeled as multi-armed bandits problem, but the classical algorithms such as upper confidence bound (UCB) algorithm are not applicable to ant colonies. This is because in ant colonies, during house-hunting, an ant can memorize only a few recently visited sites/places. In [21], we analyzed a type of learning dynamics which are well-observed in social groups like ant colonies. Specifically, under the learning dynamics of interest, an individual sequentially decides on which arm to pull next based on not only its private reward feedback but also the suggestion provided by a randomly chosen neighbor. To deal with the interplay between the randomness in the rewards and in the social interaction, we employ the *mean-field approximation* method. Considering the possibility that the individuals in the networks may not be exchangeable when the communication networks are not cliques, we go beyond the classic mean-field techniques and apply a refined version of mean-field approximation:

(i) Using coupling we showed that, if the communication graph is connected and is either regular or has doubly-stochastic degree-weighted adjacency matrix, with probability $\to 1$ as the social group size $N \to \infty$, every individual in the social group learns the best option.

(ii) If the minimum degree of the graph diverges as $N \to \infty$, over an arbitrary but given finite time horizon, the sample paths describing the opinion evolutions of the individuals are asymptotically independent. In addition, the proportions of the population with different opinions converge to the unique solution of a system of ODEs. Interestingly, the obtained system of ODEs are invariant to the structures of the communication graphs. In the solution of the obtained ODEs, the proportion of the population holding the correct opinion converges to 1 exponentially fast in time.

In addition to house-hunting, I have also worked on distributed task allocation problem in multi-agent systems, where each agent selects a task in such a way that, collectively, they achieve a proper global task allocation. Inspired by ant colonies, we proposed several scalable and efficient algorithms to dynamically allocate the agents as the task demands vary [12].

Representative Papers:

[1] Distributed Statistical Machine Learning in Adversarial Settings: Byzantine Gradient Descent (Conference: ACM SIGMETRICS 2018; Journal: ACM on Measurement and Analysis of Computing Systems, Dec. 2017) with Y. Chen and J. Xu

[2] On Learning Over-parameterized Neural Networks: A Functional Approximation Perspective (NeurIPS 2019) with P. Yang

[3] Non-Bayesian Learning in the Presence of Byzantine Adversaries (Conference: DISC 2016; Journal: Distributed Computing, Aug. 2019) with N. Vaidya

1. REFERENCES

[1] Federated learning: Collaborative machine learning without centralized training data. https://research.googleblog.com/2017/04/federated-learning-collaborative.html. Accessed: 2017-04-10.

[2] D. Alistarh, Z. Allen-Zhu, and J. Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4613–4623, 2018.

[3] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. *arXiv preprint arXiv:1807.00459*, 2018.

[4] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer. Byzantine-tolerant machine learning. *arXiv preprint arXiv:1703.02757*, 2017.

[5] Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

[6] J. Hromkovič, R. Klasing, A. Pelc, P. Ruzicka, and W. Unger. *Dissemination of information in communication networks: broadcasting, gossiping, leader election, and fault-tolerance*. Springer Science & Business Media, 2005.

[7] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

[8] A. Juditsky, A. Nazin, A. Nemirovsky, and A. Tsybakov. Algorithms of robust stochastic optimization based on mirror descent method. *arXiv preprint arXiv:1907.02707*, 2019.

[9] L. Lamport, R. Shostak, and M. Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 4(3):382–401, 1982.

[10] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[11] H. Mendes and M. Herlihy. Multidimensional approximate agreement in byzantine asynchronous systems. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing (STOC)*, pages 391–400. ACM, 2013.

[12] H.-H. Su, L. Su, A. Dornhaus, and N. Lynch. Ant-inspired dynamic task allocation via gossiping. In

International Symposium on Stabilization, Safety, and Security of Distributed Systems, pages 157–171. Springer, 2017.

[13] L. Su, C.-J. Chang, and N. Lynch. Spike-based winner-take-all computation: Fundamental limits and order-optimal circuits. *(to appear) Neural Computation*, 2019+.

[14] L. Su and S. Shahrampour. Finite-time guarantees for byzantine-resilient distributed state estimation with noisy measurements. *arXiv preprint arXiv:1810.10086*, 2018.

[15] L. Su and N. Vaidya. Multi-agent optimization in the presence of byzantine adversaries: Fundamental limits. In *2016 American Control Conference (ACC)*, pages 7183–7188. IEEE, 2016.

[16] L. Su and N. H. Vaidya. Byzantine-resilient multi-agent optimization. *https://sites.google.com/site/lilisuece/home/byzantine-resilient-multi-agent-optimization*.

[17] L. Su and N. H. Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434. ACM, 2016.

[18] L. Su and N. H. Vaidya. Non-bayesian learning in the presence of byzantine agents. In *International symposium on distributed computing*, pages 414–427. Springer, 2016.

[19] L. Su and J. Xu. Securing distributed gradient descent in high dimensional statistical learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(1):12, 2019.

[20] L. Su and P. Yang. On learning over-parameterized neural networks: A functional approximation perspective. *arXiv preprint arXiv:1905.10826*, 2019.

[21] L. Su, M. Zubeldia, and N. Lynch. Collaboratively learning the best option on graphs, using bounded local memory. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(1):11, 2019.

[22] S. Sundaram and C. N. Hadjicostis. Distributed function calculation via linear iterative strategies in the presence of malicious agents. *IEEE Transactions on Automatic Control*, 56(7):1495–1508, 2010.

[23] N. H. Vaidya and V. K. Garg. Byzantine vector consensus in complete graphs. In *Proceedings of the 2013 ACM symposium on Principles of distributed computing*, pages 65–73. ACM, 2013.

[24] C. Xie, O. Koyejo, and I. Gupta. Generalized byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116*, 2018.

[25] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett. Defending against saddle point attack in byzantine-robust distributed learning. *arXiv preprint arXiv:1806.05358*, 2018.

[26] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.