Securing Distributed Gradient Descent in High Dimensional Statistical Learning

Lili Su
Massachusetts Institute of Technology, EECS
lilisu@mit.edu

Jiaming Xu
Duke University, The Fuqua School of Business
jiaming.xu868@duke.edu

ABSTRACT

We consider unreliable distributed learning systems wherein the training data is kept confidential by external workers, and the learner has to interact closely with those workers to train a model. In particular, we assume that there exists a system adversary that can adaptively compromise some workers; the compromised workers deviate from their local designed specifications by sending out arbitrarily malicious messages.

We assume in each communication round, up to q out of the m workers suffer Byzantine faults. Each worker keeps a local sample of size n and the total sample size is $N = nm$. We propose a secured variant of the gradient descent method that can tolerate up to a constant fraction of Byzantine workers, i.e., $q/m = O(1)$. Moreover, we show the statistical estimation error of the iterates converges in $O(\log N)$ rounds to $O(\sqrt{q/N} + \sqrt{d/N})$, where d is the model dimension. As long as $q = O(d)$, our proposed algorithm achieves the optimal error rate $O(\sqrt{d/N})$. Our results are obtained under some technical assumptions. Specifically, we assume strongly-convex population risk. Nevertheless, the empirical risk (sample version) is allowed to be non-convex. The core of our method is to robustly aggregate the gradients computed by the workers based on the filtering procedure proposed by Steinhardt et al. [9]. On the technical front, 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, and show that the aggregated gradient, as a function of model parameter, converges uniformly to the true gradient function. To get a near-optimal uniform concentration bound, we develop a new matrix concentration inequality, which might be of independent interest.

KEYWORDS

Distributed systems; secured machine learning; Byzantine adversaries; high-dimensional statistics

ACM Reference Format:
Lili Su and Jiaming Xu. 2019. Securing Distributed Gradient Descent in High Dimensional Statistical Learning. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3309697.3331499

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6678-6/19/06.
https://doi.org/10.1145/3309697.3331499

1 INTRODUCTION

Distributed machine learning has been an attractive solution to large-scale problems for years. At the same time, learning in the presence of (possibly malicious) outliers has a deep root in robust statistics [7] and has become an extremely active area recently. However, most of the previous work implicitly assumes that the systems used to carry out the learning task are reliable, i.e., each computing device follows some designed specification. In this work, we consider unreliable distributed learning systems that are prone to system failures or even adversarial attacks. In particular, we assume that there exists a system adversary that can adaptively choose some computing devices to compromise; the compromised devices deviate from their local designed specifications and behave maliciously in an arbitrary manner.

Our consideration of unreliable distributed learning systems is motivated by the recent trends in a new learning framework wherein the training data is kept confidential by external computing devices, and the learner interacts with the external computing devices to train a model. In classical learning frameworks, data is collected from its providers (who may or may not be voluntary) and is stored by the learner. Such data collection immediately leads to data providers' serious privacy concerns, which root in not only purely psychological reasons but also the poor real-world practice of privacy-preserving solutions. In fact, privacy breaches occur frequently, with recent examples including Facebook data leak scandal, iCloud leaks of celebrity photos, and PRISM surveillance program. Putting this privacy risk aside, data providers often benefit from the learning outputs. For example, in medical applications, although participants may be embarrassed about their use of drugs, they might benefit from good learning outputs that can provide high-accuracy predictions of developing diseases.

To resolve this dilemma of data providers, researchers and practitioners have proposed an alternative learning framework wherein the training data is kept confidential by its providers from the learner and these providers function as workers. This framework has been implemented in practical systems such as Google's *Federated Learning* [8], wherein Google tries to learn a model with the training data kept confidential on the users' mobile devices. We refer to this new learning framework as *learning with external workers*. In contrast to the traditional learning framework under which models are trained within data-centers, in *learning with external workers* the learner faces serious *security* risk: (1) some external workers may be highly unreliable or even be malicious (hacked by the system adversary); (2) the learner lacks enough administrative power over those external workers. In this paper, we aim to develop strategies to safeguard distributed machine learning against adversarial workers while keeping the following two key practical constraints in mind:

- Small local samples versus high model dimensions: While the total volume of data over all workers may be large, individual workers may keep only small samples comparing to model dimensions. That is, the training data is *locally* a scarce resource.
- Communication constraints: Similar to other large distributed systems, the external workers are typically highly heterogeneous in terms of computation powers, real-time local computation environments, etc. As a result of this, each round of communication requires synchronization; the transmission between the external workers and the learner typically suffers from high latency and low throughout.

These two constraints together raise significant challenges for designing securing strategies. Without the first constraint, a one-shot outlier-resilient aggregation procedure suffices: each worker separately performs learning based on the local sample and sends the local estimates to the learner who aggregates these estimates to output a final global estimate. This procedure is straightforward to implement and is communication-efficient [6, 12]. However, the correctness of these algorithms crucially relies on the assumption that the local sample size is sufficiently large. In particular, $n = N/m \gg d$, where m is the number of workers, n is the local sample size, $N = nm$ is the total sample size, and d is the model dimension. In contrast, practical distributed learning systems often operate in the regime where $n \ll d$. Two immediate consequences are: (1) to learn an accurate model, the learner has to interact closely with those external workers, and such close interaction gives the adversary more chances to foil the learning process; (2) identifying the adversarial workers based on abnormality is highly challenging, as it becomes difficult to distinguish the statistical errors from the adversarial errors when the sample sizes are small. In addition, due to the randomness of the training data, the estimates computed at different rounds are highly dependent on each other.

There have been attempts to robustify stochastic gradient descent (SGD) [2, 3] with different focus from what we consider here. In particular, [3] assumes all the workers can access the whole data sample. Similar to ours, the concurrent work [2] considers the scenario where data is generated and stored in a distributed fashion at the workers. However, [2] assumes that in each iteration the workers are able to use *fresh data* to compute the gradients. However, fresh data in each round implies that the local sample size grows with time, which is not necessarily true in some applications. The fresh data assumption is crucial in their analysis: with fresh data, conditioning on the current model parameter estimator, the local gradients computed at different workers become independent, and the existing analysis of robust mean estimation may suffice. In this work, we assume that the sample size is fixed over time, and the training data is stored in a distributed fashion [5, 11].

We propose a robust gradient descent method that tolerates up to a constant fraction of adversarial workers (i.e., $\frac{q}{m} = O(1)$) and converges to a statistical estimation error $O(\sqrt{q/N} + \sqrt{d/N})$ in $O(\log N)$ communication rounds; whereas, the minimax-optimal error rate in the failure-free and centralized setting is $O(\sqrt{d/N})$. As long as $q = O(d)$, our proposed algorithm achieves the optimal error rate $O(\sqrt{d/N})$, matching the failure-free optimal error rate. Our results are obtained under some technical assumptions that we hope to relax in the future. Specifically, we assume that the population risk is strongly-convex. Nevertheless, we do allow the empirical risk (sample version) to be non-convex.

On the technical front, to deal with the interplay of the randomness of the data and the iterative updates of the model choice θ, we first establish the concentration of sample covariance matrix of gradients *uniformly* at all possible model parameters; then we prove that our aggregated gradient, as a function of θ, converges uniformly to the population gradient function $\nabla F(\cdot)$. Similar uniform concentration of sample covariance matrix has been derived in [4, Lemma 2.1] under the assumption that the gradients are sub-gaussian. While sub-gaussian *data distribution* is commonly assumed in statistical learning literature, the resulting *gradients* may be sub-exponential or even heavier tailed. Note that standard routine to bounding the spectral norm of the sample covariance matrix is available, see [10, Theorem 5.44] and [1, Corollary 3.8] for example. However, it turns out that using these existing results, the uniform concentration bound obtained is far from being optimal. To this end, we develop a new concentration inequality for matrices with i.i.d. sub-exponential column vectors. This new inequality leads to a near-optimal uniform bound.

ACKNOWLEDGMENTS

L. Su was supported in part by the NSF Science & Technology Center for Science of Information Grant CCF-0939370. J. Xu was supported in part by the NSF Grants CCF-1850743, IIS-1838124, and CCF-1856424.

REFERENCES

[1] Radosław Adamczak, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. 2010. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society* 23, 2 (2010), 535–561.

[2] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. 2018. Byzantine Stochastic Gradient Descent. *arXiv preprint arXiv:1803.08917* (2018).

[3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Byzantine-Tolerant Machine Learning. *arXiv preprint arXiv:1703.02757* (2017).

[4] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2017. Learning from Untrusted Data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2017)*. ACM, New York, NY, USA, 47–60. https://doi.org/10.1145/3055399.3055491

[5] Yudong Chen, Lili Su, and Jiaming Xu. 2017. Distributed Statistical Machine Learning in Adversarial Settings: Byzantine Gradient Descent. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 44 (Dec. 2017), 25 pages. https://doi.org/10.1145/3154503

[6] Jiashi Feng, Huan Xu, and Shie Mannor. 2014. Distributed Robust Learning. *arXiv preprint arXiv:1409.5937* (2014).

[7] Peter J Huber. 2011. Robust statistics. In *International Encyclopedia of Statistical Science*. Springer, 1248–1251.

[8] Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. 2015. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575* (2015).

[9] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 2018. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018) (Leibniz International Proceedings in Informatics (LIPIcs))*, Anna R. Karlin (Ed.), Vol. 94. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 45:1–45:21. https://doi.org/10.4230/LIPIcs.ITCS.2018.45

[10] Roman Vershynin. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* (2010).

[11] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. 2018. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. *arXiv preprint arXiv:1803.01498* (2018).

[12] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. 2013. Communication-Efficient Algorithms for Statistical Optimization. *Journal of Machine Learning Research* 14 (2013), 3321–3363. http://jmlr.org/papers/v14/zhang13b.html