Machine Learning from Distributed, Streaming Data

Waheed U. Bajwa, Volkan Cevher, Dimitris Papailiopoulos, and Anna Scaglione

I. Background

The field of machine learning has undergone radical transformations during the last decade. These transformations, which have been fueled by our ability to collect and generate tremendous volumes of training data and leverage massive amounts of low-cost computing power, have led to an explosion in research activity in the field by academic and industrial researchers. Unlike many other disciplines, advances in machine learning research are also finding rapid adoption by industry and are beginning to disrupt fields ranging from healthcare [1], journalism [2], and retail industry [3] to wireless communications [4], supply chain management [5], and automotive industry [6]. In many of the up and coming applications of machine learning in these and other fields, such as connected and/or autonomous vehicles, smart grids, edge-caching wireless networks, cloud computing, and urban policing, data are increasingly distributed and are also often streaming. Training predictive models in this distributed, streaming setting requires a rethinking of off-the-shelf machine learning solutions. A number of academic and industrial researchers have recognized the need for this in the last few years; the resulting solutions leverage algorithmic and analytical tools from a number of research areas that cut across multiple disciplines [7]–[10]. Many of these tools, such as stochastic approximation [11], [12], online learning [13], [14], distributed optimization [15], [16], and decentralized computing [17], have been the mainstay of signal processing researchers for more than a few decades. The *IEEE Signal Processing Magazine*, therefore, is one of the best forums to archive the latest advances in machine learning from data that are distributed, streaming, or both distributed and streaming, and to discuss many of the open challenges that remain to be solved for broad adoption of machine learning tools across a large number of industries that are expected to routinely deal with large volumes of distributed and/or streaming datasets.

II. An Overview of the Special Issue

This special issue on "distributed, streaming machine learning" presents recent advances in several topic areas that pertain to training of machine learning models from data that are distributed, streaming, or both distributed and streaming. A particular emphasis of articles in the special issue is to provide the reader an entry point into algorithmic and analytical techniques that may be relevant to the industry for the emerging era of real-time, decentralized, and autonomous decision making. In particular, the 13 articles comprising this special issue not only focus on potentially disruptive techniques that may form the core of future machine learning-driven systems, but they also cover techniques that are already being adopted by practitioners. These articles, authored by leading researchers in industry and academia, can be broadly categorized into seven interconnected themes within distributed, streaming machine learning, with some of the articles spanning multiple themes. These themes and their connections to the different overview articles appearing in the special issue are summarized in the following.

A. Distributed Learning

While future machine learning systems will revolve around a number of technological themes, there is one paradigm that is expected to form the core of many future systems. This paradigm, referred to as distributed learning, corresponds to an interconnected network of devices/nodes/sites in which each entity has its own set of training data and the goal is to train a global model that is as accurate as having trained it on a single machine that has access to the entire collection of data samples. This paradigm, which is already being extensively explored by academic and industrial researchers, typically arises in applications where either sharing of raw data between different entities cannot take place due to communications or privacy constraints, or where the learning task necessarily needs to be broken across multiple entities due to computational, memory, and/or storage constraints. The articles by **Nassif et al.**, **Chang et al.**, and **Cui et al.** in the special issue introduce the reader to various aspects of machine learning, which range from general convex and nonconvex learning to training of specific large-scale models for automatic speech recognition, under this distributed learning paradigm.

B. Federated Learning

The federated learning paradigm is somewhat similar to the distributed learning paradigm in that the data is still distributed across different entities. Unlike the distributed learning paradigm, however, these different entities (e.g., cell phones, wearable devices, etc.) do not communicate among themselves due to trust issues and/or communications challenges and do not transfer raw data to the cloud due to privacy concerns. Instead, in federated learning, each entity locally updates the global model using its local data and then shares the updated model with a centralized entity, which intermittently passes that model to other entities for further updates and refinements of the global model. The federated learning paradigm is increasingly gaining popularity, especially within the Web 2.0 companies, due to privacy reasons, and the article by **Li et al.** in the special issue provides an overview of the unique characteristics and challenges associated with federated learning systems.

C. Learning from Streaming Data

Streaming is another aspect of modern datasets that will occupy a central place in future machine learning systems. Indeed, many future applications of machine learning are expected to involve data sources that continuously generate data, either at a constant or at a variable rate. Streaming in conjunction with distributed datasets create additional unique challenges that require redesign of many machine learning algorithms. The articles by **Koppel et al.**, **Dall'Anese et al.**, and **Xu and Zhao** discuss myriad challenges and the corresponding solutions associated with learning from (distributed) data streams under scenarios that range from nonparametric learning and learning in dynamic environments to distributed learning in repeated unknown games.

D. Distributed Optimization for Machine Learning

Since optimization methods form the bedrock of most machine learning algorithms, distributed optimization is expected to play a major role in machine learning systems that involve distributed datasets. It is in this context that three articles in the special issue are devoted to survey of various aspects of distributed optimization that have implications for future machine learning systems. In particular, the article by **Nedić** provides an overview of distributed gradient methods for convex learning problems, the article by **Xin et al.** discusses stochastic first-order methods for distributed machine learning, and the article by **Pu et al.** explores the role of network topology in distributed stochastic optimization for machine learning.

E. Distributed Reinforcement Learning

Another subdomain of machine learning systems that will increasingly have to deal with streaming, distributed datasets is that of reinforcement learning. Roughly, the basic problem in reinforcement learning, which has a strong overlap with control theory, is to take "actions" based on observed data that maximize some notion of a "reward." Unlike control theory, however, system dynamics are assumed to be unknown in reinforcement learning and the actor/agent has to rely solely on observations for implicit "learning" of the dynamics. Distributed reinforcement learning, in which the streaming observations are also distributed, is almost certain to take center stage in so-called multiagent systems that are abstractions of applications such as autonomous vehicular networks, autonomous robot swarms, etc. The article by **Lee et al.** in the special issue provides the reader an overview of this emerging area of distributed reinforcement learning.

F. Coding Theory for Computations in Distributed Machine Learning

Practical implementations of large-scale distributed machine learning frameworks capable of handling massive datasets also require advances in coding theory for robustness against read/write (storage) errors, computation errors, component failures, communication bottlenecks, etc. In much the same way coding theory techniques enabled operation of communication systems closer to information theoretic limits, it is expected that a new generation of codes designed for distributed machine learning will enable operation of distributed processing systems closer to their theoretical limits. It is in this vein that the article by **Ramamoorthy et al.** in the special issue acquaints the reader with the use of coding theory to mitigate the effects of "stragglers," defined as slow or failed worker nodes in the system, in distributed matrix computations.

G. Distributed Adversarial Machine Learning

Given that machine learning systems are expected to be used in critical applications (e.g., management of a nation's power infrastructure and fleets of autonomous vehicles), their robustness and security against adversarial actions and malicious actors becomes paramount. While the initial focus in this direction has mostly been on centralized problems, recent works have started to develop and analyze algorithms for distributed machine learning systems that can deal with unreliable data, malicious actors, and cyber attacks on individual entities in the network. The article by **Yang et al.** in the special issue surveys recent developments pertaining to distributed adversarial machine learning under the threat model of "Byzantine attacks."

REFERENCES

[1] S. Gubar. Using A.I. to transform breast cancer care. The New York Times. Published: Oct. 24, 2019. [Online]. Available: https://nyti.ms/360AVfr

[2] S. Chandler. Reuters uses AI to prototype first ever automated video reports. Forbes. Published: Feb 7, 2020. [Online]. Available: http://bit.ly/2v92F3V

[3] J. Greig. AI, machine learning, robots, and marketing tech coming to a store near you. TechRepublic. Published: February 4, 2020. [Online]. Available: https://tek.io/2SgZDSN

[4] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[5] J. McKendrick. Artificial intelligence works its way into supply chains. ZDNet. Published: January 24, 2020. [Online]. Available: https://zd.net/38zft1D

[6] D. Lu. AI could help make fast-charging, long-lasting electric car batteries. New Scientist. Published: February 19, 2020. [Online]. Available: http://bit.ly/2vb4cXc

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, ser. Foundations and Trends in Machine Learning. Hanover, MA: Now Publishers Inc., Jan. 2011, vol. 3, no. 1.

[8] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *Proc. Conf. Neural Information Processing Systems (NeurIPS'11)*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 693–701.

[9] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *Proc. 50th Annu. Allerton Conf. Communication, Control, and Computing*, Oct. 2012, pp. 1543–1550.

[10] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server," in *Proc. 11th USENIX Symp. Operating Systems Design and Implementation (OSDI'14)*, 2014, pp. 583–598.
[11] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.
[12] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.
[13] L. Bottou, *Online learning and stochastic approximations*. Cambridge University Press, 2009.
[14] S. Shalev-Shwartz, *Online learning and online convex optimization*, ser. Foundations and Trends in Machine Learning. Now Publishers, Inc., 2012, vol. 4, no. 2.
[15] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Automat. Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.
[16] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automat. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[17] R. Bekkerman, M. Bilenko, and J. Langford, Eds., *Scaling up machine learning: Parallel and distributed approaches*. New York, NY: Cambridge University Press, 2012.

Waheed U. Bajwa (waheed.bajwa@rutgers.edu) received BE (with Honors) degree in electrical engineering from the National University of Sciences and Technology, Pakistan in 2001, and MS and PhD degrees in electrical engineering from the University of Wisconsin-Madison in 2005 and 2009, respectively. He was a postdoctoral research associate at Princeton University from 2009 to 2010 and a research scientist at Duke University from 2010 to 2011. He has been with Rutgers University since 2011, where he is currently an associate professor in the Department of Electrical and Computer Engineering and an associate member of the graduate faculty of the Department of Statistics. His research interests include statistical signal processing, high-dimensional statistics, machine learning, harmonic analysis, inverse problems, and networked systems. He has received a number of awards in his career including the Army Research Office Young Investigator Award (2014), the National Science Foundation CAREER Award (2015), and Rutgers University's Presidential Fellowship for Teaching Excellence (2017). He is a co-investigator on a work that received the Cancer Institute of New Jersey's Gallo Award for Scientific Excellence in 2017, a co-author on papers that received Best Student Paper Awards at IEEE IVMSP 2016 and IEEE CAMSAP 2017 workshops, and a Member of the Class of 2015 National Academy of Engineering Frontiers of Engineering Education Symposium. He has also been involved in a number of professional, organizational, and editorial activities. He is currently serving as a Guest Editor for a special issue of Proceedings of the IEEE, a Senior Area Editor for IEEE Signal Processing Letters, and an Associate Editor for IEEE Transactions on Signal and Information Processing over Networks. He is a Senior Member of the IEEE.

Volkan Cevher (volkan.cevher@epfl.ch) Volkan Cevher received the B.Sc. (valedictorian) in electrical engineering from Bilkent University in Ankara, Turkey, in 1999 and the Ph.D. in electrical and computer engineering from the Georgia Institute of Technology in Atlanta, GA in 2005. He was a Research Scientist with the University of Maryland, College Park from 2006-2007 and also with Rice University in Houston, TX, from 2008-2009. Currently, he is an Associate Professor at the Swiss Federal Institute of Technology Lausanne and a Faculty Fellow in the Electrical and Computer Engineering Department at Rice University. His research interests include machine learning, signal processing theory, optimization, and information theory. Dr. Cevher is an ELLIS fellow and was the recipient of the Google Faculty Research Award on Machine Learning in 2018, IEEE Signal Processing Society Best Paper Award in 2016, a Best Paper Award at CAMSAP in 2015, a Best Paper Award at SPARS in 2009, and an ERC CG in 2016 as well as an ERC StG in 2011. He is a Senior Member of the IEEE.

Dimitris Papailiopoulos (dimitris@papail.io) is an Assistant Professor of Electrical and Computer Engineering and Computer Sciences (by courtesy) at the University of Wisconsin-Madison, a faculty fellow of the Grainger Institute for Engineering, and a faculty affiliate at the Wisconsin Institute for Discovery. His research interests span machine learning, information theory, and distributed systems, with a current focus on communication-efficient training algorithms and coding-theoretic techniques that guarantee robustness during training and inference. Between 2014 and 2016, Dimitris was a postdoctoral researcher at UC Berkeley and a member of the AMPLab. Dimitris earned his Ph.D. in ECE from UT Austin in 2014, under the supervision of Alex Dimakis. In 2007 he received his ECE Diploma and in 2009 his M.Sc. degree from the Technical University of Crete, in Greece. Dimitris is a recipient of the NSF CAREER Award (2019), a Sony Faculty Innovation Award (2019), the Benjamin Smith Reynolds Award for Excellence in Teaching (2019), and the IEEE Signal Processing Society, Young Author Best Paper Award (2015). In 2018, he co-founded MLSys, a new conference that targets research at the intersection of machine learning and systems. In 2018 and 2020 he was program co-chair for MLSys, and in 2019 he co-chaired the 3rd Midwest Machine Learning Symposium.

Anna Scaglione (anna.scaglione@asu.edu; M.Sc.'95, Ph.D. '99) is currently a professor in electrical and computer engineering at Arizona State University. Her research is rooted in statistical signal processing and spans a broad number of disciplines that relate to network science, from communication, control and energy delivery systems. Her most recent work in signal processing has focused on distributed learning and data analytics for signals that are driven by network processes. Dr. Scaglione was elected an IEEE fellow in 2011. She served in many capacities, primarily the IEEE Signal Processing Society. At present she is deputy EiC for the IEEE Transactions on Control Over Networked Systems. She received the 2000 IEEE Signal Processing Transactions Best Paper and the 2013 IEEE Donald G. Fink Prize Paper Award for the best review paper in that year in the IEEE publications. Her work with her student earned several conference papers awards, and also the 2013 IEEE Signal Processing Society Young Author Best Paper Award (Lin Li). She is Distinguished Lecturer for the Signal Processing Society in 2019-2020.