RESEARCH STATEMENT Ramtin Pedarsani

Human and industrial automation, powered by artificial intelligence (AI) and the ecosystem of billions of mobile and computing devices with sensors connected through the infrastructure of the internet (i.e., the Internet of Things) is shaping the future of our society. Due to the growing computational power of these devices, future *distributed machine learning (ML) systems* operate based on storing data locally and pushing computation to edge nodes of the network. As a result, next-generation distributed learning systems will encounter a paradigm shift from the centralized setting, where data is stored and processed in a central machine, to distributed computing systems.

While distributed learning systems can enable a variety of new applications in health-care, autonomous vehicles, smart cities, robotics, etc., one needs to tackle new challenges that require a major departure from the standard methods designed for machine learning and optimization both at the systems level and algorithmic level. The core constitutive components of distributed machine learning systems are "learning and inference", "optimization", and "control". Based on this, my approach in designing efficient distributed learning systems is to a) identify *systems challenges*, develop tools from information and coding theory to tackle these challenges, and understand the *fundamental limits of computation and learning*, b) design and analyze *distributed optimization algorithms* that have provably near-optimal performance, and c) leverage the developed systematic solutions for distributed machine learning to achieve the ultimate goal of optimally controlling a multi-agent data-driven *human-cyber-physical-system* such as a next-generation intelligent transportation system. Towards this end, I bring expertise from 3 traditionally separate research areas: (i) Information and coding theory to understand the fundamental limits of high-dimensional inference, machine learning, and distributed computation; (ii) Optimization theory for designing machine learning algorithms with strong provable performance guarantees, and (iii) Control and game theory for efficient control of multi-agent networked systems and understanding humans' behavior in an autonomous system. The goal of my research group is to design theoretical foundations and algorithms that address challenges in future data-driven and autonomous systems with a major focus on distributed machine learning and computation, as well as semi-autonomous h-CPS.

Overview of Selected Work

My work has focused on several related projects on distributed learning systems, machine learning, and control. I summarize selected prior contributions in the following.

Coded Computing: Coding Theory for Fast and Reliable Distributed Computation. Modern large-scale computing systems have enabled many applications in the fields of big data analytics, mobile computing, and scientific computing, and are driven by scaling out computations across clusters consisting of many small machines made of commodity low-end hardware. There are, however, two well-known fundamental bottlenecks that arise: (1) Data Shuffling Bottleneck, which is due to the massive amounts of data that must be moved among computing nodes, often over many iterations of a running algorithm; (2) Straggler Delay Bottleneck, which is due to the latency in waiting for the slowest nodes to finish their tasks. The traditional approaches to these problems involve injecting computation redundancy in the form of "replications". For example, replicating the straggling task on another available node is a common approach to deal with stragglers. To tackle these bottlenecks, we have proposed techniques from information and coding theory to develop distributed computation schemes that are robust to stragglers and have small communication overhead. In particular, I am the co-author of the first paper that introduces how coding theoretic techniques can enable fast and reliable distributed machine learning [1]. The journal version of this paper obtained the **Communications Society & Information Theory Society joint paper award** in 2020, has received major recognition since 2016 (more than 500 citations), and led to the development of a new area known as *Coded Computing* in the information theory and machine learning communities. Since then, my group has published several conference and journal papers in the area of coded computation addressing different aspects of the problem including wireless coded computing [2], heterogeneous coded computing [3], multicore coded computing [4], coded computing for big data analytics [5], gradient coding [6], and dynamic coded computing [7]. Further, I am the co-author of the paper "Online Coded Caching" [8] that received the **best paper award at IEEE International Conference on Communications (ICC)** in 2014. This work has demonstrated significant impact in the emerging area of coded caching by introducing system dynamics and optimal cache update and eviction rules, and has resulted in a published patent [9]. The idea of using coding theory for data shuffling in machine learning is indeed closely related to online coded caching.

Robust and Communication-Efficient Federated Learning. In many large-scale machine learning applications, data is acquired and processed at the edge nodes of the network such as mobile devices, users' devices, and IoT sensors. Federated learning is a novel paradigm that aims to train a learning model at the edge nodes as opposed to traditional distributed computing systems such as data centers. The main objective of federated learning is to fit a model to data generated from network devices without continuous transfer of the massive amount of collected data from edge of the network to back-end servers for processing. Federated learning has been deployed by major technology companies including Google and Facebook with the goal of providing privacy-preserving services using users' data; however, it faces several challenges including communication bottleneck, network heterogeneity, statistical heterogeneity, privacy, and security.

My group has proposed a novel federated learning algorithm called `FedPAQ`, a communication-efficient **Fed**erated learning algorithm with **P**eriodic **A**veraging and **Q**uantization, which provably addresses the communication and scalability issues in federated learning [10] and has already received major attention from the ML community (95 citations in less than 2 years). This method stands on three key features: (1) Quantized message-passing where the edge nodes quantize their updates before uploading to the parameter server; (2) periodic averaging where models are updated locally at devices and only periodically averaged at the server; and (3) partial device participation where only a fraction of devices participate in each round of the training. Recently, we have also addressed the straggler problem using an adaptive node participation mechanism that leverages the interplay between statistical accuracy and system heterogeneity [11]. Finally, we have addressed the challenge of robustness and data heterogeneity in federated learning using optimal transport theory and federated minimax optimization in [12]. In a related line of work, we have developed the first distributed optimization algorithm with quantization that provably and exactly converges to the optimal solution at optimal rate for various settings and network topologies [13–15].

High-Dimensional Inference. The past several years have seen a new approach to the recovery of high-dimensional data, where a few *sketches* of the data retain sufficient information for an approximate sparse recovery. This approach has found numerous applications in the areas of compressive sensing, data stream computing, group testing, etc. Our group has established an innovative method for the sparse recovery of high-dimensional data using fast decoding techniques from modern coding theory. Using this method, one can solve the sparse recovery problem with strong guarantees using sublinear time complexity of $\mathcal{O}(kpolylog(n))$ where k is the sparsity and n is the data dimension, and with near-optimal sample complexity. We have applied the methodology to the problems of compressive sensing [16], compressive phase retrieval [17], group testing [18–20], and sparse covariance estimation [21]. Deploying fast decoding techniques, the proposed method significantly outperforms the sate-of-the-art sparse recovery algorithms in terms of decoding or time complexity. In a related line of work, we have studied the performance of a wide class of convex-optimization-based estimators for recovering a signal from corrupted *one-bit measurements* in high-dimensions [22]. The general result predicts *sharply* the performance of such estimators in the linear asymptotic regime using the convex Gaussian min-max theorem (CGMT) from probability theory. We have extended this result to characterize for the first time the fundamental limits on the statistical accuracy of convex ridge-regularized empirical risk minimization (ERM) for inference in high-dimensional generalized linear models [23]. For a stylized setting with Gaussian features, we have developed sharp characterization together with tight lower bounds on the estimation and prediction error that enable us to design optimal loss function and regularization parameters in supervised learning via ERM.

Adversarial Machine Learning. Developing robust training schemes against adversarial perturbations to the input data has received significant attention in the machine learning community, and is of major importance in sensitive applications such as healthcare. While deep neural networks have advanced the state-of-the-art in many machine learning applications, they are vulnerable to adversarial perturbations to the input data. As an example, one can induce image classification errors through small perturbations that are imperceptible to humans. It was initially conjectured that this vulnerability is due to the complex and nonlinear nature of the neural networks, but it was later demonstrated that it is actually due to their "excessive" linearity. Motivated by this observation, our group has focused on characterizing the fundamental limits of adversarial machine learning using stylized mathematical models. We have provided exact asymptotics for standard and adversarial errors of the estimators obtained by adversarial training with ℓ_q-norm bounded perturbations ($q \geq 1$) for both discriminative binary models and generative Gaussian mixture models [24]. We have further considered the ℓ_0 or sparse attack model, and proposed a novel asymptotically optimal classifier that is based on filtration and truncation [25]. This the first theoretical result that shows optimality of robust classification under sparse attack. We have further proposed a new robust classification method under ℓ_∞ attack based on generalized likelihood ratio test (GLRT) that has provably strong performance even in the case the the adversary's budget is not known [26]. In a related line of work, we have proposed a systematic approach to study adversarial machine learning by exploiting natural sparsity of the data and locally linear models for the network. We have proposed a defense mechanism against adversarial attacks based on a sparsifying front-end, designed to preserve desired input information while attenuating perturbations [27–29].

Human-Cyber-Physical-Systems with Mixed Autonomy. Today's society is rapidly advancing towards semi-autonomous systems that involve autonomous and human agents interacting in complex *human-cyber-physical systems* (h-CPS). As an example, autonomous and connected vehicles are soon becoming a significant part of roads normally used only by human drivers. Our group has developed new models for h-CPS and in particular traffic networks with mixed autonomy, where a fraction of cars are human-driven and the rest are autonomous [30, 31]. Under these models, we have revisited the Wardrop (Nash) equilibrium of mixed-autonomy traffic networks from a new game theoretic lens [32]. Using this insight, we have designed control algorithms for the autonomous cars' routing decisions (in case they are controllable or altruistic) such that the system reaches an optimum equilibrium [33, 34]. Moreover, we have designed optimal tolling schemes for both cases of differentiated and anonymous tolls to reduce the price of anarchy in traffic networks with mixed autonomy [35]. Our results demonstrate that with differentiated tolling, i.e. having different tolls for autonomous cars than human-driven cars, one can achieve the socially optimum equilibrium, but anonymous tolling can have small improvement. Finally, we have developed a novel approach in learning human choices of prices in autonomous transportation services versus latency, or travel time [36] via *active learning*. After Covid-19 pandemic, our group utilized this approach to investigate the effect of the pandemic on humans' travel choices. Clearly, we expected that people value safety more than latency and cost (trying to avoid public transportation or ride-sharing), and the results confirmed this. We then built on this to develop new incentive mechanisms that can guarantee safe and efficient transportation [37].

Reinforcement Learning for h-CPS. Our group has introduced a novel dynamic mixed-autonomy traffic model, and using this dynamic model, we have leveraged tools from *reinforcement learning* to dynamically and optimally route autonomous cars [38]. The key idea is quite simple yet important. If we can (partially) control the routing decisions of the autonomous cars, we can guide them such that the selfish response of humans still lead the network to an efficient equilibrium. This idea becomes more interesting in the dynamic setting, where roads can experience unexpected delays due to accidents or constructions. In response to such anomalies, a controlled fleet of autonomous cars can reroute in an online fashion to reduce the overall network congestion. Due to complicated dynamics, we have tackled this problem using reinforcement learning and shown that the near-optimal control and routing policy is indeed learned. In related works, we have studied (Semi-)Autonomous Mobility on Demand (AMoD) Systems (such as Uber) from a game theoretic viewpoint [39], and optimized their joint routing and pricing decisions using reinforcement learning [40]. At a higher level, our group is active in applying techniques from reinforcement learning to the control of mixed autonomous multiagent systems. In a recent work, we have proposed a novel training method using "gifting" for multiagent reinforcement learning in order to reach socially optimum equilibria in Markov games [41]. Gifting allows each agent to give some of their reward to other agents. We have employed a theoretical framework that captures the benefit of gifting in converging to the prosocial equilibrium by characterizing the equilibria's basins of attraction in a dynamical system and demonstrated increased convergence of high risk, general-sum coordination games to the prosocial equilibrium both via numerical analysis and experiments.

References

[1] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2017.

[2] A. Reisizadeh and R. Pedarsani, "Latency analysis of coded computation schemes over wireless networks," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1256–1263, IEEE, 2017.

[3] A. Reisizadeh, S. Prakash, R. Pedarsani, and A. S. Avestimehr, "Coded computation over heterogeneous clusters," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4227–4242, 2019.

[4] K. Lee, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Coded computation for multicore setups," in *Information Theory (ISIT), 2017 IEEE International Symposium on*, IEEE, 2016.

[5] S. Prakash, A. Reisizadeh, R. Pedarsani, and A. S. Avestimehr, "Coded computing for distributed graph analytics," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6534–6554, 2020.

[6] A. Reisizadeh, S. Prakash, R. Pedarsani, and A. S. Avestimehr, "Tree gradient coding," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2808–2812, IEEE, 2019.

[7] C.-S. Yang, R. Pedarsani, and A. S. Avestimehr, "Timely-throughput optimal coded computing over cloud networks," in *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 301–310, 2019.

[8] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, 2016.

[9] M. Maddah-Ali, U. Niesen, and R. Pedarsani, "Decentralized online cache management for digital content conveyed over shared network connections based on cache fullness and cache eviction policies," Jan. 3 2017. US Patent 9,535,837.

[10] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2021–2031, PMLR, 2020.

[11] A. Reisizadeh, I. Tziotis, H. Hassani, A. Mokhtari, and R. Pedarsani, "Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity," *arXiv preprint arXiv:2012.14453*, 2020.

[12] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," *Advances in Neural and Information Processing Systems (NeurIPS)*, 2019.

[14] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[15] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in *International Conference on Machine Learning*, pp. 9324–9333, PMLR, 2020.

[16] X. Li, D. Yin, S. Pawar, R. Pedarsani, and K. Ramchandran, "Sub-linear time support recovery for compressed sensing using sparse-graph codes," *IEEE Transactions on Information Theory*, 2019.

[17] R. Pedarsani, D. Yin, K. Lee, and K. Ramchandran, "Phasecode: Fast and efficient compressive phase retrieval based on sparse-graph codes," *IEEE Transactions on Information Theory*, 2017.

[18] K. Lee, K. Chandrasekher, R. Pedarsani, and K. Ramchandran, "Saffron: A fast, efficient, and robust framework for group testing based on sparse-graph codes," *to appear in IEEE Transactions on Signal Processing*, 2019.

[19] P. Abdalla, A. Reisizadeh, and R. Pedarsani, "Multilevel group testing via sparse-graph codes," in *Signals, Systems, and Computers, 2017 51st Asilomar Conference on*, pp. 895–899, IEEE, 2017.

[20] A. Reisizadeh, P. Abdalla, and R. Pedarsani, "Sub-linear time stochastic threshold group testing via sparse-graph codes," in *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2018.

[21] R. Pedarsani, K. Lee, and K. Ramchandran, "Sparse covariance estimation based on sparse-graph codes," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 612–619, IEEE, 2015.

[22] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Sharp asymptotics and optimal performance for inference in binary models," in *International Conference on Artificial Intelligence and Statistics*, pp. 3739–3749, PMLR, 2020.

[23] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," *International Conference on Artificial Intelligent and Statistics (AISTATIS)*, 2021.

[24] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Asymptotic behavior of adversarial training in binary classification," *arXiv preprint arXiv:2010.13275*, 2020.

[25] P. Delgosha, H. Hassani, and R. Pedarsani, "Robust classification under ℓ_0 attack for the gaussian mixture model," *arXiv preprint arXiv:2104.02189*, 2021.

[26] B. Puranik, U. Madhow, and R. Pedarsani, "Adversarially robust classification based on glrt," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3785–3789, IEEE, 2021.

[27] Z. Marzi, S. Gopalakrishnan, U. Madhow, and R. Pedarsani, "Sparsity-based defense against adversarial attacks on linear classifiers," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 31–35, IEEE, 2018.

[28] S. Gopalakrishnan, Z. Marzi, U. Madhow, and R. Pedarsani, "Combating adversarial attacks using sparse representations," *International Conference on Learning Representations (ICLR) Workshop track*, 2018.

[29] C. Bakiskan, S. Gopalakrishnan, M. Cekic, U. Madhow, and R. Pedarsani, "Polarizing front ends for robust cnns," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4257–4261, IEEE, 2020.

[30] J. Lioris, R. Pedarsani, F. Y. Tascikaraoglu, and P. Varaiya, "Platoons of connected vehicles can double throughput in urban roads," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 292–305, 2017.

[31] D. Lazar, S. Coogan, and R. Pedarsani, "Capacity modeling and routing for traffic networks with mixed autonomy," in *56th Annual Conference on Decision and Control (CDC)*, IEEE, 2017.

[32] D. Lazar, S. Coogan, and R. Pedarsani, "Routing for traffic networks with mixed autonomy," *IEEE Transactions on Automatic Control*, 2020.

[33] E. Bıyık, D. Lazar, R. Pedarsani, and D. Sadigh, "Altruistic autonomy: Beating congestion on shared roads," in *Workshop on Algorithmic Foundations of Robotics (WAFR)*, 2018.

[34] E. Biyik, D. Lazar, R. Pedarsani, and D. Sadigh, "Incentivizing efficient equilibria in traffic networks with mixed autonomy," *IEEE Transactions on Control of Network Systems*, 2021.

[35] D. A. Lazar and R. Pedarsani, "Optimal tolling for multitype mixed autonomous traffic networks," *IEEE Control Systems Letters*, 2020.

[36] E. Bıyık, D. A. Lazar, D. Sadigh, and R. Pedarsani, "The green choice: Learning and influencing human decisions on shared roads," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 347–354, IEEE, 2019.

[37] M. Beliaev, E. Bıyık, D. A. Lazar, W. Z. Wang, D. Sadigh, and R. Pedarsani, "Incentivizing routing choices for safe and efficient transportation in the face of the covid-19 pandemic," in *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, pp. 187–197, 2021.

[38] D. A. Lazar, E. Bıyık, D. Sadigh, and R. Pedarsani, "Learning how to dynamically route autonomous vehicles on shared roads," *Transportation Research Part C: Emerging Technologies (to appear)*, 2021.

[39] Q. Wei, R. Pedarsani, and S. Coogan, "Mixed autonomy in ride-sharing networks," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, pp. 1940–1950, 2020.

[40] B. Turan, R. Pedarsani, and M. Alizadeh, "Dynamic pricing and fleet management for electric autonomous mobility on demand systems," *Transportation Research Part C: Emerging Technologies*, vol. 121, p. 102829, 2020.

[41] W. Z. Wang, M. Beliaev, E. Bıyık, D. A. Lazar, R. Pedarsani, and D. Sadigh, "Emergent prosociality in multi-agent games through gifting," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.