Towards New Frontiers in Federated Learning: Objectives, Algorithms, and Real-world Challenges

Kumar Kshitij Patel

Introduction

Imagine a world where your iPhone is not just a device but an intelligent extension of you. It fine-tunes itself to your habits, needs, and preferences, proactively anticipating your requirements before you even articulate them. It functions almost as an extension of your thought process, yet remarkably, without compromising the sanctity of your privacy. This is not mere speculation but the promise of a new era in machine learning that aligns perfectly with Apple's commitment to user-centric, privacy-preserving technologies [14, 4, 11]. This transformative experience hinges on advancements in Federated Learning (FL) [24, 23, 17].

FL is more than an academic curiosity; it is an impactful operational paradigm. It has catalyzed significant scientific breakthroughs [3, 31] and has found utility across a wide range of sectors—from healthcare [20, 27], to mobile technologies [22, 26], and even the financial industry [29]. Distinct from traditional distributed optimization, FL addresses complex challenges such as decentralized data, data heterogeneity, constrained communication, partial client participation, and stringent privacy protocols with unprecedented finesse.

Over the past four years, my research has covered multiple facets of Federated Learning. I have contributed to establishing robust optimization guarantees for federated algorithms [2, 6, 12, 13, 10, 8], pioneered the design of higher-order distributed algorithms [1], and developed advanced methods for on-device personalization [7]. Additionally, I have delved into fairness and the behavior of strategic agents [4, 5] while ensuring differential privacy in complex, high-dimensional settings [11].

This research statement delineates the open questions and prevailing challenges in Federated Learning. The issues under scrutiny in this statement are categorized into three principal themes in decreasing order of abstraction: (i) pinpointing the specific mathematical objectives we aim to resolve and suitably relaxing them within the constraints inherent to FL; (ii) elucidating the information-theoretic limits of these objectives and crafting optimal algorithms that match these limits; and (iii) addressing real-world complications, such as system constraints, strategic agents, and imperfect feedback loops. Over the next two years, my research will focus on bridging these theoretical and practical gaps, laying the groundwork for rigorous academic inquiry and impactful real-world applications.

Defining and Relaxing the Mathematical Objective for Federated Learning

Research Direction A In this section, I propose the following mathematical/statistical research problems:

- A.1 Understanding rigorously how to relax and scalarize the multi-criterion optimization problem in FL to benefit from collaboration while ensuring feasible optimization.
- A.2 Characterizing *data-heterogeneity* assumptions that capture the necessary conditions for and quantify the utility of collaboration for a given client sample size.
- A.3 Moving towards a personalization-aware optimization objective: understanding how to aggregate the global (shared) and client (personal) models in a heterogeneity-aware manner.

Research Problem A.1 Most collaborative learning problems with M machines/participants can be abstracted as multi-criterion stochastic optimization problems of the following form:

$$\min_{v_1,\ldots,v_M \in \mathcal{W}} \left(F_1(v_1), \ldots, F_M(v_M)\right), \tag{P1}$$

where $F_m(v) = \mathbb{E}_{z\sim\mathcal{D}_m}[f(v;z)]$ is the objective of machine m defined using a data distribution $\mathcal{D}_m$ and a loss function f. Note that machines can solve their problems locally, i.e., without collaboration, if (i) they can fully access their objectives F_m's and (ii) they do not have computational/time restrictions. If these two assumptions hold, **P1** degenerates into M different optimization problems. However, at least one of these assumptions fails to hold in practice. For instance, assume each machine can only access a data set, $S_m \sim \mathcal{D}_m^{\otimes T}$, with $|S_m| = T$ where T is much smaller than the sample complexity to optimize F_m to some target sub-optimality ϵ. Or even in the online setting,

where at each time step, the machine m gets a sample $z_t^m \sim \mathcal{D}_m$, the time complexity to reach a good solution might be too high. As a result, using pure local training to obtain a good model can be prohibitive in the worst case and very expensive in the best case. Fortunately, in many real applications such as next-word prediction on a mobile keyboard [15], these M objectives/distributions share several similarities, and sharing information between the machines can drastically cut the total training time and sample complexity [6, 12]. This is the motivation behind federated learning that usually simplifies **P1** in two steps: first, by using a **consensus model** for all the participants,

$$\min_{v \in \mathcal{W}} \left(F_1(v), \ldots, F_M(v)\right), \tag{P2}$$

and then by **linearly scalarizing** the objective to get a simple optimization problem,

$$\min_{v \in \mathcal{W}} \frac{1}{M} \sum_{m \in [M]} F_m(v). \tag{P3}$$

These simplifications can be reasonable for some problems. However, there are some key challenges that this introduces. Unless a single model is simultaneously optimal for all the participants, it is unclear which model on the Pareto frontier of **P2** is a good choice. As such, the model obtained by optimizing the scalarization **P3** might be arbitrarily bad for some clients, thus making collaboration unfair to them. This has led to growing concerns about the utility of vanilla FL, and a series of works, including work at Apple, have looked into fairness in FL [9, 25, 28]. While there are already some works [7] that offer alternative relaxations of **P1**, we lack a concrete theoretical understanding of what these relaxations entail. A better understanding of these relaxations would require reconciling work in multi-criterion optimization with FL's real-world systems constraints and challenges [8]. This is my first proposed research problem.

Research Problem A.2 The first research problem subsumes a crucial statistical problem: measuring the "similarity" between the client distributions $\mathcal{D}_m$'s to understand for a given number of data points/time-horizon T on each device if collaboration is beneficial. Most of the existing work on data heterogeneity assumptions [19, 32][13, 10, 10] focuses on reverse-engineering assumptions that enable the analysis of optimization algorithms for problem **P3** (c.f., [8] for a survey of some of these assumptions). This can be misleading as these assumptions might be restrictive and not underline the necessary statistical conditions for the utility of collaboration. This motivates my second research problem to identify statistical conditions, such as a divergence between the data distributions, to measure the benefit of collaboration, and then suitably relax objective **P1**.

Research Problem A.3 Recall that the relaxation of objective **P1** to **P2** assumes there is a *"good-enough"* consensus model for all the agents. But unless the learning task is over-parameterized, even for problems with low data heterogeneity, this might be a limiting problem formulation. For instance, devices might benefit from collaboration for learning their shared features but not their distinct unique features. A related line of work on personalized FL offers an alternative relaxation of problem **P1**,

$$\min_{w, \theta_1, \ldots, \theta_M \in \mathcal{W}} \left(F_1(g(w, \theta_1)), \ldots, F_M(g(w, \theta_M))\right), \tag{P4}$$

where for each client, we have two models: the shared *"global model"* $w \in \mathcal{W}$ and a *"client-specific model"* $\theta_m \in \mathcal{W}$, and an aggregator function $g : \mathcal{W}^2 \to \mathcal{W}$ that specifies how to combine these models. This formulation has the advantage that we no longer need a single model which works well for all the participants. Rather, we have abstracted away the benefit of collaboration in coming up with a shared global model w. It is worth noting that this formulation can recover both objectives **P1** and **P2** for appropriately defined g. While this showcases the power of objective **P4**, how to define an aggregator function for complicated loss functions (such as ones involving neural networks) is unclear. The aggregator function not only specifies the optimization problem but also impacts the implicit bias of the final solution, i.e., which part of the Pareto frontier of **P2** is recovered. This motivates the third research problem.

Finally, it is also important to understand how the extent of data heterogeneity should impact the choice of the aggregator function. Essentially, I want to formalize the question *"how much should we personalize"*, given a problem? Note that given a good aggregation function scalarizing objective **P4** gives the following objective,

$$\min_{w, \theta_1, \ldots, \theta_M \in \mathcal{W}} \frac{1}{M} \sum_{m \in [M]} F_m(g(w, \theta_m)). \tag{P5}$$

Hanzely et al. [13] have shown that **P5** recovers several personalized federated learning (PFL), multi-task learning, and meta-learning formulations. This highlights that partial-personalization has been explored independently in several domains already. I want to reconcile this knowledge in the context of federated learning. In the next section, we move towards designing algorithms that optimize the objectives **P3** and **P5** discussed in this section.

Developing and Understanding Next Generation FL Algorithms

Research Direction B. In this section, I propose the following algorithmic research problems:

- B.1 Characterizing reasonable and sufficient data-heterogeneity assumptions and higher-order smoothness assumptions to show that local SGD can dominate trivial baselines such as large mini-batch SGD [13].
- B.2 Extending guarantees for personalized local SGD to more general function classes.
- B.3 Developing and analyzing new algorithms to identify a shared set of optima across the devices using projection or other proximal oracles implementable under communication constraints.

Research Problem B.1 Most theoretical analyses for federated optimization consider **P3** along with some local-update algorithms, the most famous of which is local SGD. Despite the above-mentioned limitations of **P3**, it can still be helpful to understand the min-max oracle and communication complexities of optimizing **P3** under low heterogeneity regimes. For instance, when the devices have shared optima, optimizing **P3** will recover them. Towards this end, there is a long line of work that focuses on understanding the convergence guarantee of local SGD in the heterogeneous setting [2, 13][30, 18, 19, 10, 32]. In a recent work [8], we show new lower bounds for local SGD as well as algorithm-independent lower bounds, that imply that under most known heterogeneity assumptions, we **can not show** that local SGD dominates decades-old baselines such as mini-batch SGD. This motivates us to continue our existing work to identify heterogeneity assumptions to show the dominance of local SGD.

Even more surprisingly, we do not have satisfactory results in the homogeneous setting when all machines share the same distribution $\mathcal{D}_m = \mathcal{D}$. On the one hand, local SGD with a single communication round is optimal for simple function classes such as convex quadratic [12], but on the other, it is dominated by simple baselines for the class of general convex functions [33]. This highlights that showing the effectiveness of local SGD requires not just understanding heterogeneity assumptions but also regularity properties such as higher-order smoothness. This idea has been explored in recent work, both ours [1] and others' [34, 10], but a complete theoretical understanding is still elusive. This motivates the first research problem I have passionately pursued in the last few years.

Research Problem B.2 Following a recent line of work [1, 21, 13, 5], we analyze a personalized variant of local SGD for solving **P5** [7]. This objective has some excellent properties that aid the convergence analysis; in particular, we expect personalized local SGD to converge even when there is no shared optimum between the objectives. For the first time for any federated optimization algorithm, we can show ***a heterogeneity-agnostic guarantee that beats all the trivial baselines*** in the strongly convex setting. Our work highlights several open questions, like understanding the general convex setting, which seems much more challenging for federated learning [8].

Research Problem B.3 Finally, in the general convex setting, which is more representative of the over-parameterized regime, there is a need to develop new algorithms beyond local SGD. Specifically, let S_m^ϵ be the ϵ sub-level set of F_m where ϵ is the target optimality. Consider the intersection of the sub-level sets of the machines, $S^\epsilon := \cap_{m\in[M]} S_m^\epsilon$. If $S^\epsilon \neq \phi$, then for **P2** we are interested in obtaining either a solution inside S^ϵ or some sub-set of S^ϵ. Recent work [16] has studied the relationship between local SGD and the projection onto convex sets (POCS) algorithm [2]. However, several open questions exist, including reconciling the algorithms for finding the intersection of convex sets into the oracle and communication model for federated learning. This is an exciting direction to understand the role of local optimization in federated learning, which can be seen as an approximate projection of the sub-level sets of the respective objectives. This motivates the third research problem.

Cross-device FL, Strategic Agents, and Sequential Decision-making

Research Direction C. In this section, I will propose the following practical research problems:

- C.1 Developing new algorithms and analysis to address the unique challenges in cross-device federated learning, such as partial client participation.
- C.2 Defining notions of distribution shift in the federated setting and providing new algorithms.
- C.3 Understanding how different patterns of client defection can impact the performance of FL algorithms in different settings: cross-silo, cross-device, and everything in between [17]—and designing incentive mechanisms that avoid or limit such strategic behavior.

Research Problem C.1 Some flavors of federated learning, such as cross-device federated learning, present unique challenges and opportunities. For instance, edge devices might have limited memory and computation power and can also participate only a few times, with some devices not participating at all. At the same time, they might have very little data and could do efficient queries on their local data to compute their global optimal solution. Thus, federated learning algorithms need to adapt to these unique considerations. A more meaningful optimization objective in this setting measures the following generalization error for some meta distribution on devices $\mathcal{P}$,

$$\min_{v \in \mathcal{W}} \mathbb{E}_{m \sim \mathcal{P}, z \sim \mathcal{D}_m}[f(v; z)]. \tag{P6}$$

The optimization algorithm for the above objective must account for multiple rounds with partial client participation. We consider this problem recently in the non-convex setting and provide almost optimal guarantees [10]. However, several open questions remain, such as removing the independent client sampling assumption.

Research Problem C.2 We recently formalized the online and bandit convex optimization setting in federated learning [9] to bridge the gap between stochastic and adaptive adversaries. But to our surprise, we find that the case with adaptive adversaries is too hard, and in the worst case, there is no benefit of collaboration with first-order feedback. We need to understand more realistic models of distribution shift in the online setting, as most deployed applications of federated learning are online problems, where assuming stationarity can be unrealistic.

Research Problem C.3 Emerging data protection laws give consumers of machine learning services the right to leave training and potentially demand their data be deleted from the machine learning model. But even in a less extreme scenario, what stops the devices from leaving training once they are content with their models? We study this question in a recent paper [5], showing how such client defections can severely harm the performance of the final consensus model. However, there are several more interesting strategic models. A combination of model-based, privacy, and monetary incentives might be required to avoid defections. A significant limitation of most existing work in this area is that it does not incorporate the multiple interactions between clients and server. This makes the strategic behavior more nuanced and requires a more clever incentive mechanism design.

Towards Truly Private Federated Learning

As this research statement ends, it is important to underscore my ongoing work on the complex issue of privacy in Federated Learning. While existing solutions encompass a range of techniques—encryption, security protocols, differential privacy, among others—an unmet need exists for a holistic framework that brings these elements together. To address this gap, I recently co-organized a workshop with participation from Apple's researchers. This event served as a platform for academic and industry researchers specializing in federated learning and differential privacy to discuss open challenges and outline a future road map. With Apple's unmatched expertise in this field, I am eager to transform these research questions into impactful, real-world solutions.

Publication List

[Own1] B. Bullins, K. Patel, O. Shamir, N. Srebro, and B. E. Woodworth. A stochastic newton algorithm for distributed convex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[Own2] A. Dieuleveut and K. K. Patel. Communication trade-offs for local-sgd with large step size. *Advances in Neural Information Processing Systems*, 32, 2019.

[Own3] M. R. Glasgow, H. Yuan, and T. Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

[Own4] N. Golrezaei, R. Niazadeh, K. K. Patel, and F. Susan. Online combinatorial optimization with fairness constraints. Under Review, 2023.

[Own5] M. Han, K. K. Patel, H. Shao, and L. Wang. On th effect of defection in federated learning and how to prevent it. Under Review, 2023.

[Own6] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

[Own7] K. K. Patel, N. Gazagnadou, L. Wang, and L. Lyu. Personalizatio mitigates the perils of local sgd for distributed heterogeneous learning. Under Review, 2023.

[Own8] K. K. Patel, M. Glasgow, L. Wang, N. Joshi, and N. Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

[Own9] K. K. Patel, L. Wang, A. Saha, and N. Srebro. Federated online and bandit convex optimization. 2023.

[Own10] K. K. Patel, L. Wang, B. Woodworth, B. Bullins, and N. Srebro. Towards optimal communication complexity in distributed non-convex optimization. In *Advances in Neural Information Processing Systems*, 2022.

[Own11] L. Wang, D. Zou, K. K. Patel, J. Wu, and N. Srebro. Private overparameterized linear regression without suffering in high dimensions. Under Review, 2023.

[Own12] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[Own13] B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Other References

[Other1] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[Other2] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

[Other3] S. E. Bergen and T. L. Petryshen. Genome-wide association studies (gwas) of schizophrenia: does bigger lead to better results? *Current opinion in psychiatry*, 25(2):76, 2012.

[Other4] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers. Protection against reconstruction and its applications in private federated learning, 2019.

[Other5] A. Bietti, C.-Y. Wei, M. Dudik, J. Langford, and S. Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In *International Conference on Machine Learning*, pages 1945–1962. PMLR, 2022.

[Other6] A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao. Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[Other7] Y. J. Cho, D. Jhunjhunwala, T. Li, V. Smith, and G. Joshi. Maximizing global model appeal in federated learning, 2023.

[Other8] Y. Collette and P. Siarry. *Multiobjective optimization: principles and case studies*. Springer Science & Business Media, 2004.

[Other9] K. Donahue and J. Kleinberg. Optimality and stability in federated learning: A game-theoretic approach, 2021.

[Other10] M. R. Glasgow, H. Yuan, and T. Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

[Other11] F. Granqvist, M. Seigel, R. van Dalen, Áine Cahill, S. Shum, and M. Paulik. Improving on-device speaker verification using federated learning with privacy. In *Interspeech*, 2020.

[Other12] N. Haghtalab, M. Jordan, and E. Zhao. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419, 2022.

[Other13] F. Hanzely, B. Zhao, and M. Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint arXiv:2102.09743*, 2021.

[Other14] K. Hao. How apple personalizes siri without hoovering up your data. *Technology Review*, 2020.

[Other15] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[Other16] D. Jhunjhunwala, S. Wang, and G. Joshi. Fedexp: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023.

[Other17] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. corr. *arXiv preprint arXiv:1912.04977*, 2019.

[Other18] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[Other19] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[Other20] W. Li, F. Milletari, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*, pages 133–141. Springer, 2019.

[Other21] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[Other22] B. McMahan and D. Ramage. Federated learning: Collaborative machine learning without centralized training data, Apr 2017.

[Other23] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016.

[Other24] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

[Other25] A. Papadaki, N. Martinez, M. Bertran, G. Sapiro, and M. Rodrigues. Minimax demographic group fairness in federated learning. In *ACM FAccT*, 2022.

[Other26] M. Paulik, M. Seigel, H. Mason, D. Telaar, J. Kluivers, R. van Dalen, C. W. Lau, L. Carlson, F. Granqvist, C. Vandevelde, et al. Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv preprint arXiv:2102.08503*, 2021.

[Other27] K. Powell. Nvidia clara federated learning to deliver ai to hospitals while protecting patient data. *Nvidia Blog*, 2019.

[Other28] B. Rodriguez-Galvez, F. Granqvist, R. van Dalen, and M. Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS Workshop*, 2021.

[Other29] G. Shiffman, J. Zarate, N. Deshpande, R. Yeluri, and P. Peiravi. Federated learning through revolutionary technology " consilient, Feb 2021.

[Other30] S. U. Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[Other31] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[Other32] J. Wang, R. Das, G. Joshi, S. Kale, Z. Xu, and T. Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

[Other33] B. E. Woodworth, B. Bullins, O. Shamir, and N. Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.

[Other34] H. Yuan and T. Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.