Research Statement

Kumar Kshitij Patel kkpatel.ttic.edu

1 Introduction: Transforming AI through Multi-Distribution Learning

Humans have an exceptional ability to quickly learn new tasks by recognizing patterns and integrating prior knowledge through shared representations in the brain [66, 30, 8, 18, 20, 9, 59]. Since the early days of AI, researchers have aimed to replicate this multi-task learning capacity [55, 42, 45, 56, 65, 12, 3, 53, 22], yet most recent breakthroughs have come from improving learning performance on single tasks[1] by scaling up models, datasets, and computation [10, 61, 31]. We are now at a critical juncture where further scaling faces multiple challenges. Acquiring high-quality data is increasingly difficult due to privacy regulations and intellectual property concerns [21, 67, 2]. Worse still, collecting large datasets in fields like healthcare and drug discovery remains infeasible altogether [15, 16, 50]. Moreover, the resources needed to operate at scale—amidst diminishing returns—have concentrated power within a few companies, slowing down the pace of research [7, 4, 1, 74, 5, 64, 37, 23]. Compounding this, current AI systems continue to struggle with bias, lack robustness, and fail to generalize under distribution shifts [4, 57, 62, 25, 63].

My research in **Federated Learning (FL)** addresses many of these challenges. FL uses decentralized training across agents[2], which preserves data privacy, tackles regulatory constraints, and supports learning across multiple data distributions, thus leveraging the strengths of multi-task learning[3] [41, 40, 28]. Over the past five years, I have contributed to various aspects of FL—and more broadly **multi-distribution learning**—including fairness [3, 8], personalization [13], privacy [15, 14], sequential decision-making [11], strategic agent behavior [4], and communication-efficient optimization [2, 7, 16, 17, 1, 12, 9, 10]. I have also helped further our understanding of robustness [6] and continual learning under distribution shifts [5], both crucial for current AI algorithms. Our work has won accolades such as the **Distinguished Paper Award** at IJCAI 2024 and the **Best Paper Honorable Mention Award** at the Federated Learning Workshop at ICML 2023 while inspiring a plethora of follow-up studies. Beyond research, I have actively shared developments in the field through a **tutorial at UAI 2023** and fostered academic and industry collaborations by co-organizing a **2023 workshop on FL and privacy**.

A central theme of my research is understanding, ***how learning one task helps another***—a question at the heart of multi-distribution learning. To address this, I combine theoretical frameworks like min-max complexity with empirical insights to develop efficient algorithms that adapt to data heterogeneity. While FL has already enabled unprecedented collaboration across many fields [19, 49, 52, 68, 13, 36, 29, 58, 39, 47], expanding its adoption further requires more scalable, robust, and provably secure solutions that respect the needs of diverse applications and align with agents' incentives. As illustrated in Figure 3, achieving this requires interdisciplinary solutions. My vision is to develop such solutions that redefine AI's role as a force for meaningful societal impact.

Outline. This statement presents my vision for advancing multi-distribution learning over the next five years. To frame this vision, I introduce a taxonomy of the field in Figures 1 and 2, mapping its connections to related areas and situating my contributions within a broader context. While real-world applications present a diverse array of challenges, I focus on three key issues: first, in Section 2, I address learning with known tasks or distributions (A and C in Figure 1); second, in Section 3, I examine fine-tuning large foundation models in dynamic, multi-user environments (H and I in Figure 2); and finally, in Section 4, I explore privacy and strategic agent behavior which are critical challenges to sustainable data markets in AI (c.f., Figure 3).

[1] There are notable exceptions like OpenAI's Whisper [51] that do benefit from multi-task learning. [3] In FL, different agents share the same objective but differ in their distributions, making learning simpler than usual multi-task learning. Depending on the application, we use tasks, agents, and distributions interchangeably.

2 The Role of Collaboration and Personalization: a Tale of Three Regimes

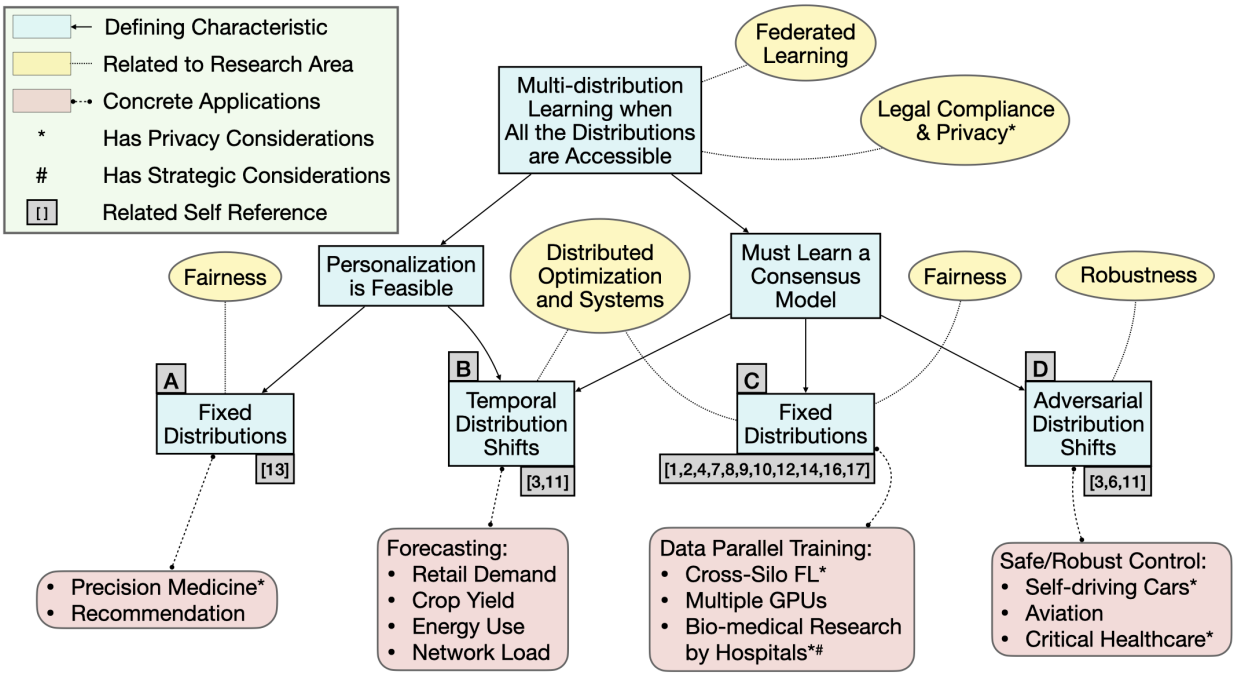

Figure 1: A taxonomy of problems with multiple distributions when **all** distributions are accessible during the learning phase. The figure highlights connections to different research areas, real-world applications, and my contribution to these different problem classes (A-D); c.f. my publications.

The most studied setting in multi-distribution learning is where distributions are fixed in time and accessible during training (say, via sampling). Assuming there are M such distributions, $\{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$, using a loss function[4] $f(\cdot;\cdot)$, we define the m^{th} objective as $F_m(v) := \mathbb{E}_{z\sim\mathcal{D}_m}[f(v;z)]$. Our aim then is to optimize the following problem for given thresholds $\{\tau_1, \ldots, \tau_M\}$,

$$\begin{aligned} \min_{w\in\mathbb{R}^d} \quad & \frac{1}{M}\sum_{m\in[M]} F_m(w) \\ \text{s.t.} \quad & F_m(w) - \min_{w_m^\star\in\mathbb{R}^d} F_m(w_m^\star) \le \tau_m, \ \forall\ m\in[M] \end{aligned} \tag{C}$$

There are several ways to motivate (**C**). When $f(\cdot;z)$ is convex, then different thresholds recover the multi-objective optimization problem with M objectives [17][13]. Another perspective is to look at the extremes of (**C**): (i) when τ_m's are set to ∞, (**C**) recovers the FL objective [40] while (ii) when τ_m's are set to $\tau^\star = \min_{w\in\mathbb{R}^d}\max_{m\in[M]} F_m(w)$, (**C**) recovers the group distributional robust objective [54, 44]. Thus, (**C**) interpolates between the **utilitarian** and **egalitarian** extremes.

Our results for optimizing (C). In a recent work [3], we propose a general framework for solving problem (**C**), given access to oracles that can (approximately) optimize a single objective. While our work in [3] focused on solving NP-hard combinatorial problems in offline and online settings, I have also extensively contributed to the design and analysis of learning algorithms for the two extreme cases (i) and (ii) in problem (**C**). In particular, I have helped characterize optimal convergence rates for Local-SGD—the most widely-used method in federated learning ($\tau_m = \infty$)—and min-max complexities for various problem classes using *intermittently communicating* algorithms[5] [2, 16, 17, 1, 12, 9, 10]. Our work has been impactful, closing long-standing gaps in these areas that had

[4] E.g., in supervised linear regression, given feature-label pairs $\boldsymbol{z} = (\boldsymbol{x}, y) \sim D_m$ if we use square loss then we can define $f(\boldsymbol{v};\boldsymbol{z}) := \frac{1}{2}(\langle \boldsymbol{v}, \boldsymbol{x}\rangle - y)^2$. [5] These are algorithms where agents communicate their oracle responses every K time steps for R rounds, e.g., mini-batch and local SGD.

persisted for over a decade [38, 73, 72, 60, 34, 35, 33, 71, 24, 70]. In a recent study [8], we also examined the robust extreme ($\tau_m = \tau^\star$), achieving the **tightest** convergence guarantees for linear regression using a second-order algorithm when M is large. Moving forward, I plan to extend these works on both extremes, addressing existing gaps and adapting to additional applications, but more importantly, I want to develop general-purpose and efficient algorithms to directly optimize (**C**) for machine learning tasks with arbitrary thresholds.

Introducing personalization and characterizing three regimes. Solving (**C**) can lead to undesirable outcomes when the heterogeneity between different distributions is high [69, 70] [9, 10]. A ubiquitous way to avoid this while still benefiting from solving multiple tasks is by using two models: a shared model $w \in \mathcal{W}$ and a task-specific model $\theta_m \in \mathcal{C}$, and then combining them using an aggregation function $g : \mathcal{W} \times \mathcal{C} \to \mathbb{R}^d$ [43, 48] which leads to the following problem,

$$\begin{aligned} \min_{w \in \mathcal{W}, \theta_1, \ldots, \theta_M \in \mathcal{C}} \quad & \frac{1}{M} \sum_{m \in [M]} F_m \left(g(w, \theta_m) \right) \\ \text{s.t.} \quad & F_m \left(g(w, \theta_m) \right) - \min_{w_m^\star \in \mathbb{R}^d} F_m \left(w_m^\star \right) \leq \tau_m, \ \forall m \in [M] \end{aligned} \tag{A}$$

The solutions to (**A**) are constrained by the choices of $\mathcal{W}$, $\mathcal{C}$, g, and the optimization algorithm used. A simple choice is $g(w, \theta) = w + \theta$, representing **additive personalization** [26, 6]. In an upcoming work [13], we analyze this additive model with $\tau_n = \infty$ and show that a personalized variant of local SGD outperforms it under equivalent computation and communication budgets. Moving forward, I aim to generalize these results to arbitrary thresholds and aggregation functions and explore personalization through a **min-max complexity** framework[6]. To achieve this, we compare two algorithm classes: consensus algorithms producing a single model and personalized algorithms yielding M models. We use non-collaborative learning—where each agent optimizes on its own—as a natural baseline. Given constraints on information or computation, and communication, I want to characterize three distinct regimes based on data heterogeneity and min-max optimality:

I. **Very high heterogeneity**: when non-collaborative learning is at least as good as any personalized or consensus optimization algorithm;

II. **Moderate heterogeneity**: when some personalized algorithm is strictly better than any consensus optimization algorithm as well as non-collaborative learning; and

III. **Low heterogeneity**: when some consensus optimization algorithm is at least as good as any personalized algorithm and strictly better than non-collaborative learning.

Theorem-of-alternatives. While these regimes appear self-evident, some problems may not realize all three of them. For instance, we prove that regime **II** does not exist when estimating multiple Gaussian means, meaning either not collaborating or consensus optimization is optimal, with no benefit from intricate personalization [13]. Although our result is for the worst case, understanding these regimes for different problems can significantly inform the practice and design of algorithms. Going forward, I aim to study problems beyond mean estimation while incorporating other factors, such as noise levels, sample sizes, and optimization complexity, into these regimes.

3 Optimizing for Fine-tuning on Unseen Tasks, a.k.a. Learning to Learn

In the previous section, we assumed all distributions would be available during training. However, this assumption may be too restrictive for many real-world applications, such as training AI models across billions of smartphones. To model such scenarios, we assume a meta-distribution $\mathcal{P}$ over

[6] Min-max complexity identifies optimal algorithm performance against worst-case distributions from some family.

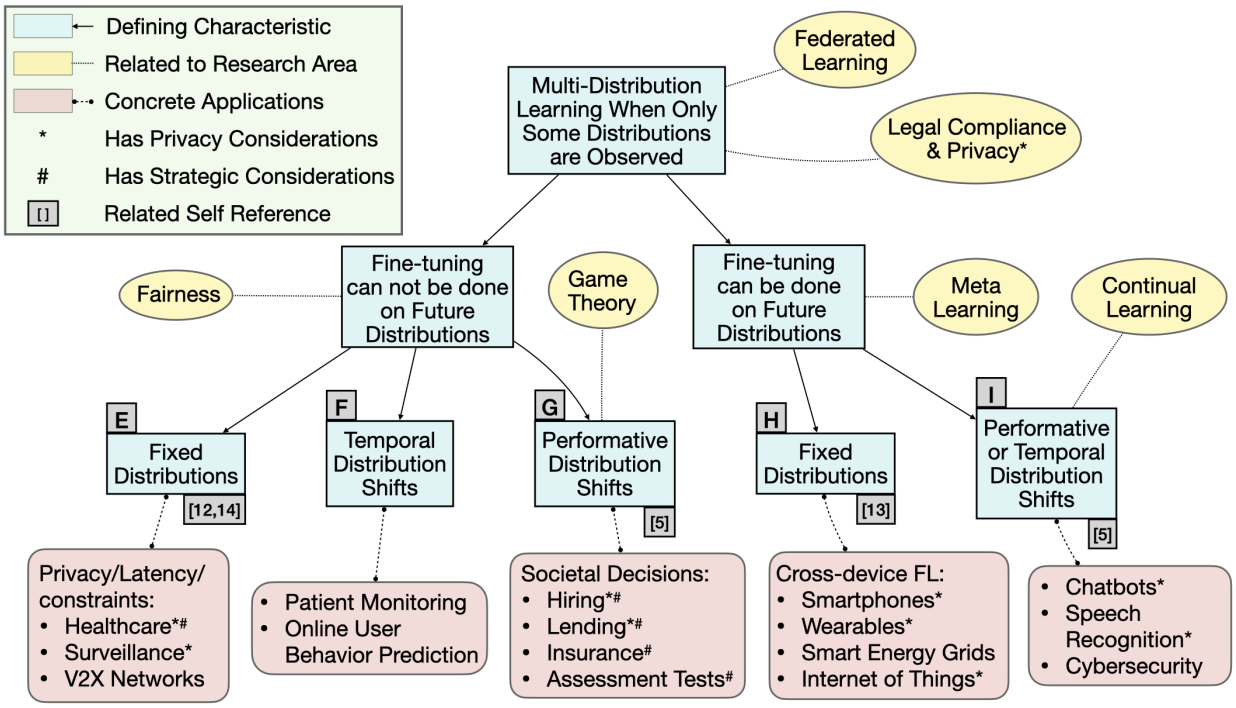

Figure 2: A taxonomy of problems when only **some** distributions are accessible during the learning phase through sampling from a meta-distribution. The figure highlights connections to different research areas, real-world applications, and my contribution to these different problem classes (E, G, H, I); c.f. my publications.

tasks, where each task $m \sim \mathcal{P}$ is associated with a distribution $\mathcal{D}_m$. If $\mathcal{E}$ is a fine-tuning algorithm that uses N samples from $\mathcal{D}_m$ to produce an output $\theta_m^{\mathcal{E}} \in \mathcal{C}$, then we aim to optimize the following,

$$\min_{w \in \mathcal{W}} \mathbb{E}_{m \sim P, z \sim D_m, \mathcal{E}} \left[f(g(w, \theta_m^{\mathcal{E}}); z) \right] \quad , \tag{\textbf{H}}$$

where $g : \mathcal{W} \times \mathcal{C} \to \mathbb{R}^d$ is an aggregator function. One common fine-tuning procedure $\mathcal{E}$ is to run N steps of SGD on θ for a model parameterized by $g(w, \theta)$. Most existing theory for cross-device FL examines a simpler form of (**H**) where $N = 0$, meaning no fine-tuning occurs on future tasks. This no-fine-tuning variant of (**H**) aligns with Problem E in Figure 2 [32] [12, 14]. In practice, however, models are often fine-tuned directly on client devices, revealing a gap between the theory and practice of FL. Although a few theoretical works study fine-tuning in this setting, they do not provide provable benefits of pre-training over random initialization in practical regimes [14, 46]. To address this gap, I plan to investigate assumptions about the heterogeneity of the meta-distribution $\mathcal{P}$, such as characterizing its fat tail or solutions that are approximately optimal across tasks. Specifically, I aim to characterize how small N must be, as a function of the pre-training budget T[7] and the heterogeneity of $\mathcal{P}$, for pre-training to outperform random initialization consistently.

Learning to learn. The effectiveness of pre-training also depends on what access to $\mathcal{P}$ is allowed during pre-training. A natural approach involves sampling $[M] \sim \mathcal{P}^{\otimes M}$ and solving either problem (**C**) or (**A**) across these M sampled tasks to produce a pre-trained model. This approach suggests that the considerations about task heterogeneity from the previous section remain relevant and become even more nuanced when we incorporate fine-tuning in problem (**H**). When data heterogeneity is high, and N is small, solving (**A**) during the pre-training stage should be preferable. Conversely, with more similar tasks or more data, (**C**) may be more effective. However, understanding the exact trade-off between these approaches, N, T, and data heterogeneity would require analyzing the min-max complexity of optimizing (**H**), which I am interested in pursuing. Notably,

[7] Intuitively, this threshold N should decrease as T increases.

solving (**A**) during pre-training resembles meta-learning approaches, as we explicitly encourage the shared model to capture the information needed for efficient fine-tuning by using the same aggregation function g and constraint set $\mathcal{C}$ during both pre-training and fine-tuning. This means that insights from studying (**H**) could also address gaps in the theory of meta-learning.

Towards continual learning. In many applications, we can sample from the meta-distribution $\mathcal{P}$ multiple times during training or continuously improve the AI model between deployments. This makes adjusting for shifts in $\mathcal{P}$ over time essential. These shifts may follow predictable patterns; for instance, when training AI models on smartphones, the distribution $\mathcal{P}$ often cycles between different states since phones typically only train when left charging overnight [69, 27]. Additionally, performative effects may emerge when the data adapts in response to model updates. For example, using human feedback to train large foundation models can introduce new errors, so avoiding "forgetting" prior corrections during retraining becomes essential. In recent work, we investigated this challenge and showed that even simple problems benefit from regularization during retraining to reduce performative effects [5]. Although many heuristics address these issues, we still lack a systematic understanding of robust algorithmic foundations for continual learning. This motivates me to explore how general regularized retraining procedures can effectively leverage multi-task learning (or fine-tuning) to handle distribution shifts.

Pre-training informed fine-tuning. Finally, I aim to improve the efficiency of the fine-tuning algorithm $\mathcal{E}$. One of the most widely-used approaches, LORA [26], accomplishes this by encoding a low-rank structure in the fine-tuned parameters. Our research on low-rank dynamics in neural network optimization [18, 19] suggests that aligning the initialization of LORA's parameters with the low-rank structure in the pre-trained weights could make fine-tuning even faster.

4 Towards Creating Sustainable Data Markets for the Future of AI

So far, we have treated agents and tasks as interchangeable, overlooking that agents often act strategically in the real world. For example, hospitals that collaborate on training AI models face privacy regulations and competitive pressures. As a result, hospitals hesitate to collaborate unless they are sure it is *individually rational*, meaning the benefits outweigh the costs. A potential solution is to use (**A**) or (**C**) with a threshold set at $\tau_n = \epsilon_n$, which represents the error an agent can achieve alone. If this maintains feasibility, it ensures individual rationality, providing further motivation to study these problems with general thresholds.

Preventing defections. Another related issue arises when agents agree to collaborate but leave during multi-stage training, abandoning the collaboration once an intermediate model satisfies their needs. In a recent work [4], we showed that such defections can significantly compromise the final model's accuracy and robustness, particularly in applications like medical studies, where models are often used on future agents without additional fine-tuning. Our work also highlighted that popular federated learning algorithms like local SGD fail to prevent defections, and we introduced **the first algorithm** that avoids defections while converging to an optimum that is shared across agents. I aim to extend this work to scenarios where agents may misrepresent their data or requirements, with the ultimate goal being an incentive-compatible collaboration mechanism that ensures agents truthfully share information and provide quality training updates. This will involve exploring monetary incentives, data valuation and attribution, and developing a deeper understanding of legal and economic considerations about compensating updates to a model (c.f., Figure 3).

Provable Privacy at Scale. Privacy remains a critical barrier to data markets, driving my research on rigorous theoretical guarantees for foundational privacy problems and large-scale, practical challenges in complex models. Recently, we adapted one of our FL algorithms to incorporate formal differential privacy (DP) guarantees in the shuffled model [12, 14], which removes the need for a trusted server, thus broadening DP's applicability. I aim to develop more such DP variants that protect against specific attack models relevant to applications in Figures 1 and 2.

We are also working towards methods that selectively apply differential privacy (DP), protecting only the model's sensitive components while relaxing constraints elsewhere to improve efficiency. This selective approach is critical for training large models, such as diffusion models, which are prone to memorizing training data and currently face challenges in effective differentially private training [11]. To address these issues, we are exploring two directions. First, inspired by our work on personalization [13], we aim to partition the diffusion model into shared and personal parameters for early and later de-noising stages, respectively. This approach protects agents' data privacy while benefiting from shared parameters in the challenging early de-noising phase and is especially relevant in medical imaging, where data is siloed across hospitals. Second, we are developing architecture-aware techniques for adding DP noise to diffusion models by identifying components prone to memorization so that the entire model can be released. This draws inspiration from our work for over-parameterized linear regression, where structural insights allow for privacy even in infinite dimensions [15]. In the future, I want to advance these efforts to models beyond diffusion models and tackle other real-world privacy challenges.

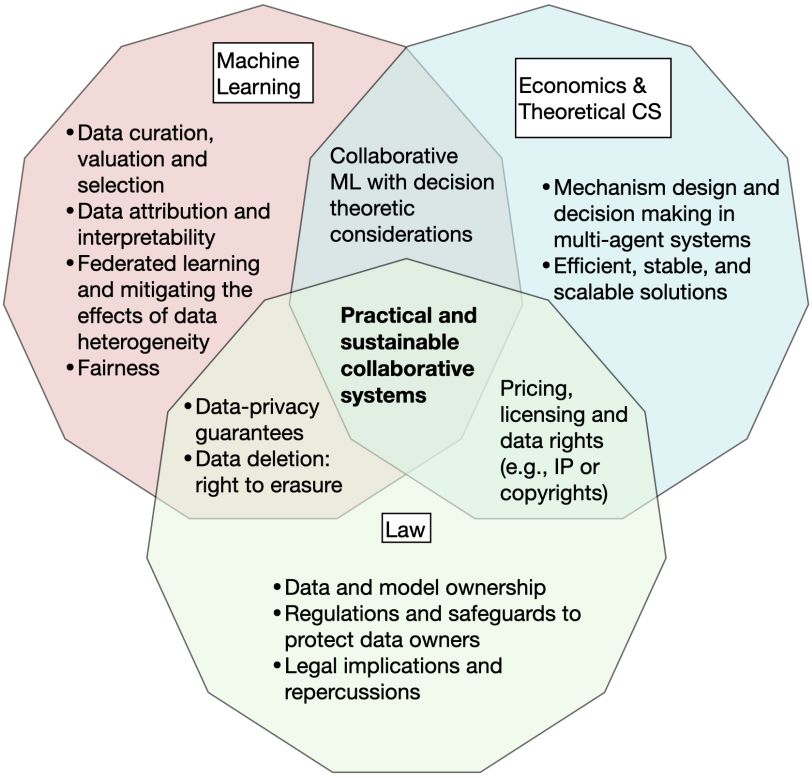

Figure 3: Key research challenges for creating and sustaining data markets, as studied across multiple disciplines.

My Publication List

[1] B. Bullins, K. 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.

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

[3] N. Golrezaei, R. Niazadeh, K. K. Patel, and F. Susan. Online combinatorial optimization with group fairness constraints. *33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[4] M. Han, K. K. Patel, H. Shao, and L. Wang. On the effect of defections in federated learning and how to prevent them. *arXiv preprint arXiv:2311.16459*, 2023. Under review.

[5] A. Kabra and K. K. Patel. The limitations of model retraining in the face of performativity. *Humans, Algorithmic Decision-Making and Society: Modeling Interactions and Impact, Workshop at International Conference on Machine Learning (ICML)*, 2024.

[6] S. Kapoor, K. K. Patel, and P. Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

[7] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local sgd. In *International Conference on Learning Representations*, 2019.

[8] N. Manoj and K. K. Patel. A second-order algorithm for empirical group distributionally robust regression. *16th International OPT Workshop on Optimization for Machine Learning at Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[9] 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.

[10] K. K. Patel, M. Glasgow, A. Zindari, L. Wang, S. U. Stich, Z. Cheng, N. Joshi, and N. Srebro. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. In S. Agrawal and A. Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4115–4157. PMLR, 30 Jun–03 Jul 2024.

[11] K. K. Patel, L. Wang, A. Saha, and N. Srebro. Federated online and bandit convex optimization. In *International Conference on Machine Learning*, pages 27439–27460. PMLR, 2023.

[12] 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.

[13] K. K. Patel, B. Woodworth, N. Gazagnadou, L. Wang, and L. Lyu. Personalization mitigates the perils of local sgd for distributed heterogeneous learning. Ongoing work, 2024.

[14] L. Wang, X. Zhou, K. K. Patel, L. Tang, and A. Saha. Efficient private federated non-convex optimization with shuffled model. In *Privacy Regulation and Protection in Machine Learning Workshop*, 2024.

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

[16] 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.

[17] 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.

[18] D. Yunis, K. K. Patel, P. H. P. Savarese, G. Vardi, J. Frankle, M. Walter, K. Livescu, and M. Maire. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[19] D. Yunis, K. K. Patel, S. Wheeler, P. Savarese, G. Vardi, K. Livescu, M. Maire, and M. R. Walter. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*, 2024. Under review.

Other References

[1] N. Ahmed and M. Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.

[2] G. Appel, J. Neelbauer, and D. A. Schweidel. Generative ai has an intellectual property problem. *Harvard Business Review*, 7, 2023.

[3] J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[5] T. Besiroglu, S. A. Bergerson, A. Michael, L. Heim, X. Luo, and N. Thompson. The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv preprint arXiv:2401.02452*, 2024.

[6] 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.

[7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[8] M. Botvinick, A. Weinstein, A. Solway, and A. Barto. Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current opinion in behavioral sciences*, 5:71–77, 2015.

[9] M. M. Botvinick and J. D. Cohen. The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive science*, 38(6):1249–1285, 2014.

[10] T. B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[11] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[12] R. Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[13] S. Chen, D. Xue, G. Chuai, Q. Yang, and Q. Liu. Fl-qsar: a federated learning-based qsar prototype for collaborative drug discovery. *Bioinformatics*, 36(22-23):5492–5498, 2020.

[14] G. Cheng, K. Chadha, and J. Duchi. Fine-tuning is fine in federated learning. *arXiv preprint arXiv:2108.07313*, 3, 2021.

[15] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15(141):20170387, 2018.

[16] M. Clark. Machine learning needs big data to revolutionise drug discovery. *Drug Discovery World*, 2021. Accessed: 2024-11-10.

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

[18] A. G. Collins and M. J. Frank. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190, 2013.

[19] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.

[20] J. Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179, 2010.

[21] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), May 2016. Legislative Body: OP_DATPRO.

[22] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[23] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer. Ai and memory wall. *IEEE Micro*, 2024.

[24] 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.

[25] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[27] D. Jhunjhunwala, P. Sharma, A. Nagarkatti, and G. Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916. PMLR, 2022.

[28] 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.

[29] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima Jr, J. Mancuso, F. Jungmann, M.-M. Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021.

[30] N. Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170, 2010.

[31] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[32] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020.

[33] 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.

[34] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[35] 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.

[36] 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.

[37] A. J. Lohn and M. Musser. Ai and compute: How much longer can computing and progress, drive artificial intelligence. 2022.

[38] R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009.

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

[40] 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.

[41] 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.

[42] M. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

[43] K. Mishchenko, R. Islamov, E. Gorbunov, and S. Horváth. Partially personalized federated learning: Breaking the curse of data heterogeneity. *arXiv preprint arXiv:2305.18285*, 2023.

[44] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

[45] A. Newell. Human problem solving. *Upper Saddle River/Prentive Hall*, 1972.

[46] J. Oh, S. Kim, and S.-Y. Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.

[47] 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.

[48] K. Pillutla, K. Malik, A.-R. Mohamed, M. Rabbat, M. Sanjabi, and L. Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pages 17716–17758. PMLR, 2022.

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

[50] A. Prakash. Exploring new chemical space for the treatments of tomorrow. *American Pharmaceutical Review*, 2023. Accessed: 2024-11-10.

[51] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[52] H. R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B. C. Bizzo, et al. Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, pages 181–191. Springer, 2020.

[53] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[54] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[55] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[56] J. Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[57] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

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

[59] B. Sorscher, S. Ganguli, and H. Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119, 2022.

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

[61] R. Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[62] C. Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[63] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[64] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10, 2020.

[65] S. Thrun. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, 8, 1995.

[66] M. Tomasello. *Primate cognition*. Oxford University Press, 1997.

[67] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2022.

[68] 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.

[69] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[70] 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.

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

[72] J. Zhang, C. De Sa, I. Mitliagkas, and C. Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.

[73] M. Zinkevich, M. Weimer, L. Li, and A. Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

[74] S. Zuboff. The age of surveillance capitalism: The fight for a human future at the new frontier of power, edn. *PublicAffairs, New York*, 2019.