Research Statement

Kumar Kshitij Patel kkpatel.ttic.edu

Scaling has driven much of the recent progress in machine learning: larger models, more computation, and ever-larger datasets [73, 54]. Yet access to the data needed to sustain this progress is increasingly uncertain, as the era of abundant, lightly regulated web data draws to a close amid copyright, consent, and privacy constraints [76, 49, 42]. At the same time, consequential applications in medicine and scientific discovery depend on scarce, proprietary, or expensive-to-collect data, from rare-disease cohorts to drug-screening results [35, 33, 75]. These data are often siloed across institutions, with heterogeneous populations and modalities. Further progress, therefore, cannot rely on scaling alone. It will also require methods that combine information across sources [69, 53].

My research develops efficient machine-learning methods for learning across heterogeneous, decentralized, and sensitive sources. Across these problems, I use theory to explain practical failures, identify exploitable structure, and design better algorithms. This agenda has four directions:

§ 1 optimizing efficiently at scale;
§ 2 adapting objectives, estimators, and models to heterogeneity;
§ 3 learning and generating from sensitive data under privacy constraints; and
§ 4 valuing data and accounting for strategic participation.

1 Efficient Optimization at Scale: From Federated Learning to LLM Training

The Unreasonable Effectiveness of Local Updates. Mini-batch SGD has been the workhorse of large-scale machine learning for two decades: workers compute stochastic gradients in parallel and average them to obtain low-variance gradient estimates [36, 37, 45]. At larger scales, however, communication increasingly dominates runtime and energy costs, making more computation per round valuable [64, 53]. Historically, this led to *large-batch training*, using extra computation to average more gradients and reduce variance; *Local SGD*—or Federated Averaging—instead uses the same budget for several sequential local updates before averaging models [45, 64]. Under heterogeneous objectives, these updates can make models drift apart, turning extra work into a source of error. Local SGD has nevertheless proved highly successful in practice [6], underpinning federated learning across millions of devices [78, 53] and inspiring data-center LLM training methods [39]. For years, however, theory explained these gains only when client drift was tightly controlled. Early analyses, including ours, assumed homogeneous data, where workers sample from the same distribution [2, 16][71], while extensions to heterogeneous data required client gradients to remain uniformly similar [15][57]. Yet Local SGD remains effective with highly dissimilar gradients, leaving its success under substantial heterogeneity "unreasonable" [79].

The prevailing view attributed this gap to overly stringent theory and expected Local SGD to require only that clients approximately share an optimizer. We showed that this weaker condition alone cannot guarantee an advantage over mini-batch SGD [8] and an early version of this work received a **Best Student Paper Honorable Mention** at the ICML Federated Learning Workshop [24]. Concurrently, related distributed methods offered a different clue: similarity in client curvature, rather than gradients, can sharply reduce communication complexity [70, 55, 72, 59, 50, 51]. Complementing this, our non-convex analysis established near-minimax-optimal communication rates under this structure for local update methods [13]. These clues led us to study curvature similarity in Local SGD. We first established its benefits for Local SGD for strongly convex objectives [14] and then for general convex objectives [9], controlling gradient disagreement along the average model's trajectory rather than globally. We also provide nearly matching lower bounds that isolate when these gains are fundamental [8, 9]. Together, these results resolve a central question in the decade-long study of min-max optimal convergence rates for Local SGD that identify when curvature similarity enables gains over mini-batching [63, 84, 83, 71, 57, 58, 56, 82, 80, 44, 79, 61].

Beyond First-Order Local Updates. Local SGD uses extra computation between communication rounds via repeated first-order updates, but richer local information can enable more powerful local computation. In our stochastic Newton method, clients use Hessian-vector products and Local SGD to approximately solve a Newton subproblem, performing one Newton update per communication round and improving communication complexity compared to first-order methods [1].

Feedback and the Value of Collaboration. A different question arises when objectives evolve over time: before asking how to communicate efficiently, we should ask whether collaboration helps at all. In federated online optimization, we showed that the answer can depend fundamentally on the feedback available to each client. Observing gradients can make collaboration unnecessary, whereas observing only incurred losses can make it beneficial in high dimensions [12]. Ongoing work further sharpens these guarantees [19]. We further developed this interface between online and federated learning in our UAI tutorial, *Online Optimization Meets Federated Learning* [31].

Structure-Aware Optimization for LLMs. Modern optimizers increasingly exploit matrix structure and geometry rather than treating weights coordinatewise [62, 47, 60, 77, 52]. In ongoing work, we go beyond matrix structure by designing updates around the computations the weights perform. We use the probabilistic geometry of transformer attention to derive mirror-descent updates with a scalable Gauss–Newton approximation [25]. On a 353M-parameter GPT-2 Medium–scale model, our attention-aware hybrid reaches intermediate validation-loss targets using up to 16% less measured training time than Newton–Muon [40], with gains concentrated early. Together with our spectral-dynamics work [23], these results motivate optimization that adapts to both architecture and training stage.

2 Adapting Learning to Heterogeneous Populations

Heterogeneous populations create two distinct challenges. When a shared model is appropriate, gains for one group may come at another's expense, so we must choose and optimize the right trade-off efficiently. When it is not, we must decide what can still be shared and what should remain local, including which components to personalize and to what extent. This broader agenda motivated our Simons workshop on *Learning from Heterogeneous Sources* [30].

Choosing and Optimizing Shared Objectives. Our group-fairness framework traces the utility–fairness Pareto frontier for general optimization problems, even computationally hard ones, in offline and online settings, with access only to an approximation algorithm for a single objective [3]. This work received the **IJCAI Distinguished Paper Award**. We also ask how efficiently such trade-offs can be optimized in structured problems. For linear regression, block Lewis-weight geometry improves worst-group optimization at moderate accuracy: the number of linear-system solves adapts to the smaller of the data rank and number of groups, while the formulation spans average to worst-group loss [7]. This raises a broader question: can such structure-adaptive, high-accuracy iteration bounds extend to general convex quadratic programs, as they do for linear programming? A complementary challenge arises when the trade-off itself evolves with the model: in LLM post-training, we study how one model can accommodate heterogeneous preferences as its updates change which populations provide feedback [68, 67].

Choosing What to Share. The sharing question arises at both the estimator and model levels. When clinical sites recruit different populations, pooling records can bias estimates for the target population. We developed distributed estimators that correct site-specific selection bias using only shared summary statistics, while achieving nearly the same statistical efficiency as if all records were pooled in one place [5]. At the model level, our personalized Local SGD method combines a shared model with client-specific corrections, interpolating between consensus and isolated training without

gradient-similarity assumptions [21]. Ongoing work extends this to reusable shared components, separating cluster-level structure from local variation.

3 Exploiting Structure for Private Learning

Formal privacy creates a utility–privacy trade-off, with stronger guarantees often requiring more data or reducing the quality of learned models or generated samples. My work asks how problem structure can mitigate these costs without weakening formal guarantees.

Differentially Private Diffusion Models. Diffusion models have transformed image generation, but their tendency to memorize training examples makes them particularly risky in sensitive domains [34]. Training them from scratch with formal differential privacy also remains difficult [38, 43]. We exploit diffusion's progressive noising and refinement: clients share only clipped, forward-noised examples, while local denoisers recover fine-grained detail. This yields, to our knowledge, the **first collaborative diffusion framework with per-record local differential privacy guarantees** for personalized image generation [10], improving generation especially for underrepresented classes without exposing clean data. Our follow-up moves collaboration into latent space to handle high-resolution images [18]. We show that privacy–utility gains depend on the distributional structure of the latent representation, not dimensionality reduction alone.

Structure-Aware Private Optimization. Beyond generative modeling, I study how structure can make private optimization less costly, a theme also reflected in the TTIC workshop *New Frontiers in Federated Learning*, which I co-organized [28]. In overparameterized linear regression, we show that privacy cost can depend on data geometry rather than parameter count; public unlabeled data can reveal this geometry and guide how noise is added [22]. Gradient clipping poses a different challenge: it limits sensitivity but can bias optimization away from the true solution. When sample losses share a common minimizer, we show that clipping bias can vanish. This yields DP-SGD guarantees without the restrictive uniformly bounded-gradient assumption, addressing a longstanding gap in private optimization [20][32, 41]. The results extend to approximate common minimizers and to minibatching, with lower bounds that identify when these guarantees are tight. Finally, with distributed data, curvature similarity also improves convergence rates for private nonconvex optimization without a trusted server, paralleling our non-private results [27][13].

4 Economics of Collaborative Learning

Theoretical Foundations of Data Valuation. Deciding which data to acquire, share, or pay for requires measuring their value for learning. Score-based methods such as Shapley values, leave-one-out scores, and influence functions support pricing, attribution, pruning, and federated client selection, but their popularity has outpaced theory on when they guide good decisions. We bridge this gap by connecting data valuation to submodular maximization and asking when selecting the highest-scoring points yields a near-optimal subset. This viewpoint identifies three structural properties—redundancy, complementarity, and non-monotonicity—that govern when scalar scores succeed or fail. Redundancy can make Shapley and leave-one-out selection arbitrarily poor even for monotone submodular valuations. Bounded curvature controls this effect and yields tight guarantees for both methods. When points become valuable only in combination with others—that is, when they complement one another—both fixed scores and greedy selection can fail. We show that sufficient dataset coverage can control this effect by restoring approximate submodularity. Finally, when adding data can be harmful, nonadaptive selection can fundamentally fail compared to adaptive methods [11].

Private valuation introduces another economic tension. A seller must reveal enough for a potential buyer to assess usefulness without giving away the benefit of purchasing the data. We

study a formal version of this tension through two-party differentially private testing, proving lower bounds and interactive protocols for assessing whether two private datasets jointly admit an accurate classifier while protecting each party's records throughout the interaction.

Strategic Collaboration. Collaboration can also fail when participants stop following the intended protocol. My earlier work studied learning under adversarially corrupted feedback [4]; more recently, I have focused on strategic deviations driven by participants' incentives. One such failure mode is defection: a client may participate honestly, then leave once an intermediate model meets its needs. We show that such departures can harm population accuracy and fairness, that Local SGD cannot generally prevent them, and that our alternative method avoids harmful defections under a shared-optimum condition [17]. A client may instead choose to remain but misreport its objective or updates. Our initial work shows that leave-one-out-based procedures can make truthful reporting optimal in structured settings [26]. Ongoing extensions combine these ideas and auditing to seek protocols that make deviations either detectable or too weak to selectively harm other clients. These questions also motivated our TTIC workshop on *Incentives for Collaborative Learning and Data Sharing* [29].

5 Future Research Directions

Architecture-Aware Optimization for Foundation Models. As frontier training spans hundreds of thousands of accelerators [46, 81, 48], communication and memory increasingly constrain how models can be trained. At this scale, energy and water use further raise the stakes for efficiency. Building on §1, I will develop a theory of architecture-aware optimization that predicts how update geometry should adapt across model components, training stages, and scale. Rather than treating optimizer design as a collection of benchmark-specific recipes, I want to understand when richer structural and curvature information is worth its cost and how those choices interact with distributed execution. The long-term goal is a set of predictive principles for the compute–quality trade-offs of foundation-model training.

Adaptive Collaboration under Heterogeneity. The objective and sharing choices in §2 point to a broader goal: learning systems should infer when and how to collaborate rather than assuming the collaboration structure in advance. I will develop a statistical and algorithmic theory spanning global models, clustered collaboration, personalized sharing, and independent learning, quantifying the cost of discovering useful partners and shared structure under limited data and communication. As populations, data, and participant needs evolve, collaborations should be able to form, split, and reorganize accordingly. Ultimately, I want collaboration itself to become a learned and adaptive part of the system.

Preserving Rare Events and Diverse Preferences. Across §2–§3, a recurring failure of average-case learning is that important events or populations can disappear from the observed data because they are rare or become harder to observe. I will develop objectives and guarantees that preserve such behavior rather than allowing it to be absorbed by the dominant distribution. For generative models, this means understanding and improving tail fidelity; for preference learning, it means actively acquiring feedback when deployment changes who continues to interact with the model. The long-term goal is to distinguish failures of representation from failures of observation and build systems that improve poorly served populations even when the evidence needed to do so is scarce or endogenous.

Auditable and Incentive-Compatible Collaboration. Building on §4, I will develop foundations for collaboration that remains reliable when participants are strategic. This requires connecting three goals—detecting deviations, limiting their harm, and making honest participation

optimal—under realistic communication, validation, and computational constraints. I will characterize the statistical and computational limits of auditing multiple adaptive or colluding clients, including how privacy can simultaneously obscure manipulation and limit its influence. The goal is a theory of the resources needed for collaboration that is simultaneously useful, private, auditable, and strategically stable.

Private Learning from Longitudinal Electronic Health Records. Healthcare generates enormous amounts of data, yet a 2019 estimate cited by the World Economic Forum suggests that 97% of hospital data globally goes unused [65]. Policy has long sought both standardized exchange and privacy protection, from HIPAA to the 21st Century Cures Act [74, 66], yet major technical barriers remain: hospitals serve different populations, outcomes unfold over time, rare cases matter disproportionately, and patient-level privacy limits what can be pooled. Building on our work on multi-site bias correction and private generation (§2–§3), I will develop a unified theory and methodology for inference and generation from longitudinal records across institutions without centralizing them. This requires jointly handling selection bias, temporal dependence, heterogeneous populations, and formal patient-level privacy rather than solving each in isolation. I will develop shared–local generative models and privacy mechanisms that account for where clinically meaningful and identifying information concentrates across time, modalities, and pathology. The long-term goal is to make siloed clinical records a usable foundation for multi-institutional discovery while preserving patient privacy and institutional control.

Refereed Conference and Journal Publications

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

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

[3] N. Golrezaei, R. Niazadeh, K. K. Patel, and F. Susan. Online combinatorial optimization with group fairness constraints. In *International Joint Conference on Artificial Intelligence*, 2024. Distinguished Paper Award.

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

[5] R. Kundu, M. Salvatore, K. K. Patel, L. Ohno-Machado, H. Cho, X. Shi, and B. Mukherjee. Privacy-enhancing sequential learning under heterogeneous selection bias in multi-site electronic health records data. *Journal of the American Medical Informatics Association*, 2026.

[6] 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*, 2020.

[7] N. S. Manoj and K. K. Patel. Distributionally robust linear regression with block Lewis weights. In *International Conference on Learning Representations*, 2026. Earlier version presented at the OPT Workshop at NeurIPS 2024.

[8] 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 *Conference on Learning Theory*, volume 247, pages 4115–4157. PMLR, 2024.

[9] K. K. Patel, R. Islamov, S. U. Stich, A. Lucchi, E. Gorbunov, and L. Wang. What's in a smoothness constant? tighter rates for local SGD with bounded second-order heterogeneity. In *Advances in Neural Information Processing Systems*, 2026.

[10] K. K. Patel, B. Jiang, A. F. M. M. Kabir, W. Zhang, D. Zou, and L. Wang. Personalized federated training of diffusion models with privacy guarantees. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[11] K. K. Patel, S. P. Karimireddy, R. C. Fernandez, and M. Zampetakis. Limits and potential of score-based data valuation: Redundancy, complementarity, and non-monotonicity. In *Advances in Neural Information Processing Systems*, 2026. Earlier version presented at the DATA-FM Workshop at ICLR 2026.

[12] K. K. Patel, L. Wang, A. Saha, and N. Srebro. Federated online and bandit convex optimization. In *International Conference on Machine Learning*, 2023.

[13] 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.

[14] K. K. Patel, A. Zindari, S. U. Stich, and L. Wang. Revisiting consensus error: A fine-grained analysis of local SGD under second-order data heterogeneity. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[15] B. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems*, 2020.

[16] B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, 2020.

Manuscripts Under Review or In Submission

[17] 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. In submission.

[18] A. F. M. M. Kabir, K. K. Patel, D. Zou, and L. Wang. Privacy-preserving personalized federated latent diffusion models. *Manuscript under review*, 2026.

[19] K. K. Patel et al. Sharper theory for federated online and bandit optimization. Ongoing work; manuscript in preparation, 2026.

[20] K. K. Patel, A. E. Gourari, and A. Koloskova. Differentially private sgd without the bounded gradients assumption: The boon of interpolation, 2026. Manuscript under review.

Public Preprints

[21] K. K. Patel, N. Gazagnadou, L. Wang, and L. Lyu. Personalization mitigates the perils of local SGD for heterogeneous distributed learning. Preprint, 2023.

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

[23] 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. Preprint, arXiv:2408.11804, 2024.

Refereed Workshop Papers and Extended Abstracts

[24] 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 Workshop at ICML*, 2023. Best Student Paper Honorable Mention.

[25] K. K. Patel and N. S. Manoj. Attention geometry helps early: Mirror-descent updates for transformer query and key matrices. In *OPT Workshop at NeurIPS*, 2026.

[26] K. K. Patel and A. Murhekar. Performative personalization incentivizes truthfulness in federated learning. In *AIMS Workshop at ICLR*, 2026.

[27] 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 at ICLR*, 2024.

Workshops and Tutorials Organized

[28] K. K. Patel. New frontiers in federated learning. Workshop at TTIC, 2023. Co-organizer.

[29] K. K. Patel. Incentives for collaborative learning and data sharing. Workshop at TTIC, 2025. Co-organizer.

[30] K. K. Patel. Learning from heterogeneous sources. Workshop at the Simons Institute, 2026. Co-organizer; designed and led the open-problems session.

[31] K. K. Patel and A. Saha. Online optimization meets federated learning. Tutorial at UAI, 2023.

Other References

[32] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security*, 2016.

[33] J. Banerjee, J. N. Taroni, R. J. Allaway, D. V. Prasad, J. Guinney, C. S. Greene, et al. Machine learning in rare disease. *Nature Methods*, 20:803–814, 2023.

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

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

[36] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, 2012.

[37] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(6):165–202, 2012.

[38] T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

[39] A. Douillard, Q. Feng, A. A. Rusu, R. Chhaparia, Y. Donchev, A. Kuncoro, M. Ranzato, A. Szlam, and J. Shen. DiLoCo: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

[40] Z. Du and W. Su. The Newton–Muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

[41] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 2006.

[42] 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 (general data protection regulation), May 2016.

[43] S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

[44] 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*, 2022.

[45] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[46] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[47] V. Gupta, T. Koren, and Y. Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 2018.

[48] S. Guthrie. Inside the world's most powerful AI datacenter, Sept. 2025. The Official Microsoft Blog, September 18, 2025.

[49] P. Henderson, X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

[50] X. Jiang, A. Rodomanov, and S. U. Stich. Federated optimization with doubly regularized drift correction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21912–21945. PMLR, 21–27 Jul 2024.

[51] X. Jiang, A. Rodomanov, and S. U. Stich. Stabilized proximal-point methods for federated optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[52] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. Technical writeup.

[53] 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. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

[54] 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.

[55] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. Breaking the centralized barrier for cross-device federated learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[56] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

[57] A. Khaled, K. Mishchenko, and P. Richtarik. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[58] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

[59] D. Kovalev, A. Beznosikov, E. D. Borodich, A. Gasnikov, and G. Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. In *Advances in Neural Information Processing Systems*, 2022.

[60] H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, 2024.

[61] P. Mangold, A. O. Durmus, A. Dieuleveut, S. Samsonov, and E. Moulines. Refined analysis of constant step size federated averaging and federated Richardson–Romberg extrapolation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5023–5031. PMLR, 2025.

[62] J. Martens and R. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015.

[63] R. McDonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, volume 22, pages 1231–1239, 2009.

[64] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[65] J. Moore and Y. D. Guichot. How to harness the power of health data to improve patient outcomes. World Economic Forum, Jan. 2024. January 5, 2024.

[66] Office of the National Coordinator for Health Information Technology. 21st century cures act: Interoperability, information blocking, and the ONC health IT certification program. Final Rule, 85 FR 25642, May 2020. Published May 1, 2020.

[67] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, 2020.

[68] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

[69] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020.

[70] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[71] S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

[72] Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, 2022.

[73] R. S. Sutton. The bitter lesson. Incomplete Ideas, 2019.

[74] U.S. Department of Health and Human Services. General overview of standards for privacy of individually identifiable health information. HIPAA Privacy Rule Guidance, 2015. Last reviewed November 5, 2015.

[75] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18:463–477, 2019.

[76] 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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49523–49544. PMLR, 2024.

[77] N. Vyas, D. Morwani, R. Zhao, I. Shapira, D. Brandfonbrener, L. Janson, and S. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

[78] 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.

[79] J. Wang, R. Das, G. Joshi, S. Kale, Z. Xu, and T. Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *Transactions on Machine Learning Research*, 2024.

[80] B. E. Woodworth, B. Bullins, O. Shamir, and N. Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4386–4437. PMLR, 2021.

[81] xAI. Grok 4, July 2025. xAI, July 9, 2025.

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

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

[84] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 23, pages 2595–2603, 2010.